Anymotion: A Dataset, Benchmark, and Baseline for Controllable Human Motion Editing

Haiyang Yan[1,2,*], Jianxin Sun[3,*], Yuhan Wu[3,4,*], Libin Wang[3,†], Zhiwei Li[1], Man Zhang[4], Jingdong Chen[3], and DanDan Zheng[3,†]

[1]Institute of Automation, Chinese Academy of Sciences

[2]School of Future Technology, University of Chinese Academy of Sciences

[3]Ant Group [4]Beijing University of Posts and Telecommunications

Abstract

Human motion change is currently hindered by the lack of large-scale datasets, comprehensive benchmarks, and robust algorithms capable of handling diverse motion requirements. To address these gaps, we propose AnyMotion, a novel open-world paradigm for highly controllable and identity-consistent human motion editing across arbitrary scenes and actions. Our contributions are two-fold. First, we establish a hierarchical label taxonomy covering six major categories and 527 fine-grained classes, and curate HME-1M, a large-scale dataset containing 1.04 million editing quadruplets via systematic data balancing, cleaning, and reverse construction. Second, we introduce the Skeleton Chain-of-Thought (SCoT) framework, a two-stage pipeline consisting of motion reasoning and guided generation. In the reasoning stage, a multimodal large language model serves as a cognitive planner to derive target skeletal keypoints and MetaQueries. These outputs subsequently function as explicit structural and semantic priors, providing the necessary guidance for the diffusion transformer to perform high-fidelity motion generation. Finally, built upon our label taxonomy, we present Human Motion-Bench, the first motion-centric benchmark equipped with multi-category fine-grained evaluation metrics. Extensive experiments demonstrate that AnyMotion achieves state-of-the-art performance across diverse editing scenarios.

The project is available at `https://github.com/Haiyang0226/AnyMotion`.

1 Introduction

Human motion change has broad application potential in digital content creation [1], virtual avatar control [2], and film post-production [3]. However, early formulations [4–6] of this task are largely video-driven and typically assume static backgrounds and temporally continuous motions. Such assumptions restrict models to small, smooth motion transitions and make them ill-suited for non-contiguous, cross-scene motion editing, which significantly diverges from the high-degree-of-freedom editing demands observed in open environments such as social media. Recent advances in multimodal large language models (MLLMs) [7–9], particularly in fine-grained instruction understanding and high-fidelity generation, provide a key opportunity to overcome these constraints. Motivated by this progress, we redefine the human motion change task as an open-world paradigm that aims to edit human motions while preserving identity consistency, enabling "any scene, any action" editing.

*Contributed equally to this work. †Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite its importance, progress on human motion change has been fundamentally constrained by the lack of high-quality training data and systematic evaluation benchmarks. From the data perspective, existing methods [10–12] typically construct training samples by sampling adjacent frames from real or generated videos. These samples mainly capture small and continuous displacements under nearly static backgrounds, making them inadequate for modeling large pose variations or scene changes required by realistic editing scenarios. Meanwhile, current mainstream benchmarks [6, 13, 11] contain extremely limited samples related to human motion change, with narrow motion categories and small pose variations. Furthermore, there is still no fine-grained evaluation protocol that assesses generation quality across different motion categories. The combination of biased data distributions and incomplete evaluation dimensions substantially hinders progress toward highly flexible and controllable motion editing.

To fill the gap of large-scale real-scene human motion editing data, we construct Human Motion Edit-1M (HME-1M), a dataset containing 1.04 million high-quality samples. We first establish a hierarchical annotation taxonomy covering 527 motion categories. Through scene- and pose-aware clustering, we perform balanced down-sampling for dominant categories and targeted augmentation for underrepresented ones to alleviate the long-tail distribution. We then treat real images as target motion images and employ multimodal large models to inversely construct the corresponding source images and editing instructions, followed by strict quality control for the generated editing pairs. To enable comprehensive evaluation, we further introduce Human Motion-bench based on the same taxonomy, which contains real human images from diverse scenes and editing instructions spanning rich motion categories. We also design a fine-grained evaluation protocol that measures both overall image quality and motion accuracy conditioned on motion categories.

To enable accurate and highly controllable target motion generation, we introduce the Skeleton Chain-of-Thought (SCoT) paradigm. Departing from conventional single-step "instruction-to-pixel" mappings, SCoT adopts a "think-before-generate" strategy that decouples high-level intent reasoning from low-level visual synthesis. Specifically, we fine-tune the MLLM to predict target skeletal keypoints alongside implicit MetaQuery features [14]. The explicit keypoints establish a precise structural prior, effectively bridging the gap between natural language instructions and concrete human poses, while the MetaQueries encode contextual semantics. These dual signals are jointly injected into a Diffusion Transformer (DiT) backbone as synergistic conditioning, ensuring anatomically plausible and instruction-faithful synthesis. Furthermore, the framework supports user-provided skeleton inputs during inference, enabling precise, fine-grained customization of pose and motion trajectories. In summary, our contributions are as follows:

- We transcend the conventional limits of motion editing long constrained by video-driven temporal continuity and static backgrounds and introduce AnyMotion, a new openworld paradigm that empowers freeform motion generation for any scene, any action.
- For the first time, we establish a comprehensive pipeline covering data collection, construction, cleaning, and benchmarking for human motion change, delivering HME-1M (1.04M high-quality samples with a hierarchical motion taxonomy) and Human Motion-benchthe first benchmark with fine-grained metrics across motion categories, thereby filling the long-standing gaps in both data resources and evaluation.
- We propose a SCoT paradigm that performs accurate motion reasoning and leverages semantic and structural guidance for controllable motion editing. Extensive experiments demonstrate that our approach achieves state-of-the-art performance, establishing a new standard for highly controllable, open-world human motion editing.

2 Related Works

2.1 Human Motion Change

Image editing [15–18] has evolved from early appearance-level attribute manipulation toward higher-level control over spatial structure and motion [19–23]. Among these tasks, human motion editing aims to precisely modify the pose or action of a human subject according to textual instructions while strictly preserving identity consistency. However, progress in this area remains constrained by limitations in high-quality data construction, evaluation protocols, and methodologies.

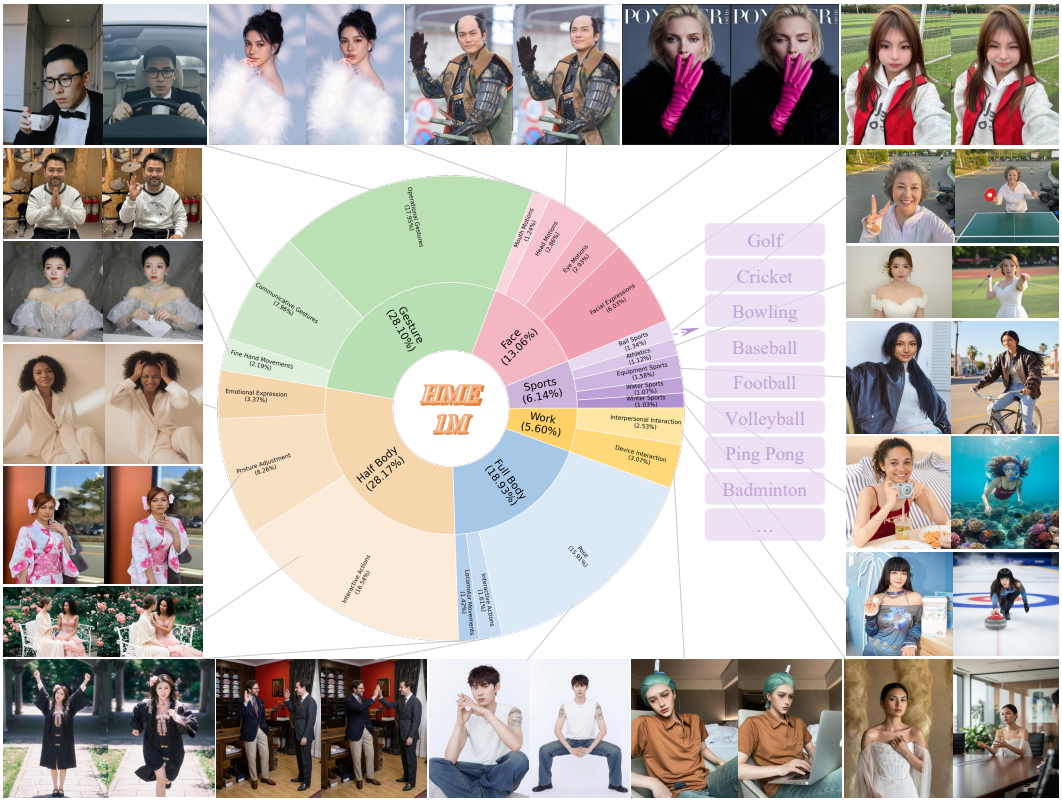

Figure 1: Label taxonomy and HME-1M dataset. The label taxonomy is organized into four levels and covers 527 actions across six major categories. Based on this taxonomy, we construct HME-1M, which contains 1.04M representative real images.

Recent studies construct motion editing datasets from video sequences. InstructMove [12] and Step1X-Edit [6] rely on optical flow estimation and foreground segmentation to filter video frames, extracting samples that exhibit physically plausible motion and clear subject displacement. MotionEdit [11] further expands the data source to high-quality generated videos, leveraging MLLMs for motion semantic verification and instruction refinement. Additionally, Frame2Frame [10] utilizes image-to-video models to construct smooth temporal transition paths on the image manifold, implicitly maintaining content consistency before and after editing. Despite these advances, such video-driven paradigms are inherently constrained by the temporal continuity of videos, making it difficult to cover real-world scenarios that involve cross-scene, large-scale, or discontinuous motion changes.

Meanwhile, existing image editing datasets and evaluation benchmarks (e.g., GEdit [6] and GEdit-v2 [13]) primarily emphasize instruction adherence, overall visual quality of the generated image, and visual consistency in non-edited regions. Motion change is typically treated as only one of many generic editing sub-tasks, lacking dedicated evaluation criteria for the physical plausibility of motion, as well as the accuracy of motion magnitude and direction. MotionEdit [11] is the first to explicitly formulate motion image editing as an independent task and introduces a motion alignment score based on optical flow alignment. However, the human motion categories it covers remain limited, and it has not yet established fine-grained, structured evaluation standards tailored to different types of actions (e.g., full-body pose, facial expressions, or specialized sports movements), making it difficult to precisely assess the dynamic generation fidelity of editing models.

2.2 Pose-Guided Image Generation

Pose-controlled image generation aims to synthesize target images with high visual fidelity and anatomically plausible human structures conditioned on explicit pose representations. With the rise of latent diffusion models (LDMs) [24], U-Netbased architectures [25–27], exemplified by Control-

Net [19] and T2I-Adapter [23], have become the dominant paradigm. However, this framework is inherently limited in modeling long-range dependencies, often leading to anatomical inconsistencies such as limb disconnections or joint misalignments under extreme poses or multi-person scenarios. Recently, Diffusion Transformers (DiT) [28] have demonstrated stronger capabilities in conditional generation due to their global self-attention mechanisms and scalable tokenized representations, significantly improving structural coherence and conditioning flexibility. Representative works such as HumanDiT [29] encode pose conditions into structured tokens and inject them into the denoising process via cross-attention or adaptive layer normalization, effectively capturing long-range dependencies within human body topology. Nevertheless, these approaches largely rely on direct mappings from static pose representations and lack explicit reasoning or planning for complex editing instructions and underlying physical dynamics. As a result, they struggle to bridge the semantickinematic gap between high-level intent and low-level joint motion.

To address this limitation, we propose a hierarchical editing framework following a "think-before-generate" paradigm. This design decouples intent reasoning from pixel-level generation, enabling coordinated optimization between faithful instruction adherence and natural integration with environmental context in complex human pose editing scenarios.

3 Dataset and Benchmark

In this section, we focus on constructing a large-scale, high-quality image editing dataset for human motion change that is realistic and covers diverse scenes and action types. Section 3.1 introduces the construction of the human motion change label taxonomy. Section 3.2 describes the dataset construction pipeline. Section 3.3 presents the proposed HumanMotion Bench.

3.1 Label Taxonomy

We collect 1.77M high-quality real human motion images from the web and filter out samples with ambiguous or unclear action categories. We first employ Gemini 3.0 to annotate the human motions in these images and merge labels with highly similar semantics. The resulting label sets are then organized through hierarchical semantic clustering into a four-level taxonomy that covers six major categories: full-body actions, upper-body actions, facial expressions, hand gestures, sports, and actions related to work.

From the web-sourced images, we observe a noticeable lack of professional sports and occupational action labels within the initial taxonomy. We therefore supplement these categories with targeted actions, bringing the combined set of initial and supplementary labels to 840. After manual merging and deduplication, we obtain a final set of **527 fine-grained fourth-level labels** that span a wide range of everyday and professional human motions, as illustrated in Figure 1.

The distribution of web-collected data exhibits a long-tail pattern, where some motion categories contain significantly more samples than others. Directly using such data for training may cause models to overfit to dominant classes. To address this issue, for categories containing more than 20k images, we perform clustering based on scene context and human pose features, followed by balanced sampling to retain 20k images per category while preserving diversity. For categories with fewer than 200 images, we synthesize additional data using the generative model Gemini 3.0, resulting in 90k newly generated images. This process yields a more balanced distribution across action categories.

3.2 Dataset Construction Pipeline.

To enable motion-controllable image editing, we construct quadruples consisting of a source image, editing instruction, target pose skeleton, and target image, as

$$(I_{src}, T_{edit}, \mathcal{K}_{tgt}, I_{tgt}). \tag{1}$$

As shown in Figure 2, we use real images as target images, which ensures high-quality supervision, and detect their keypoints; we then reversely construct the corresponding source image and editing instruction. To generate diverse source images with large pose variations while avoiding logically inconsistent motion transitions, we first use Qwen3-VL-8B [30] to generate a reverse editing instruction and the corresponding forward instruction for each image. Conditioned on these instructions, Gemini 3.0 [31] is then used to synthesize the corresponding source image.

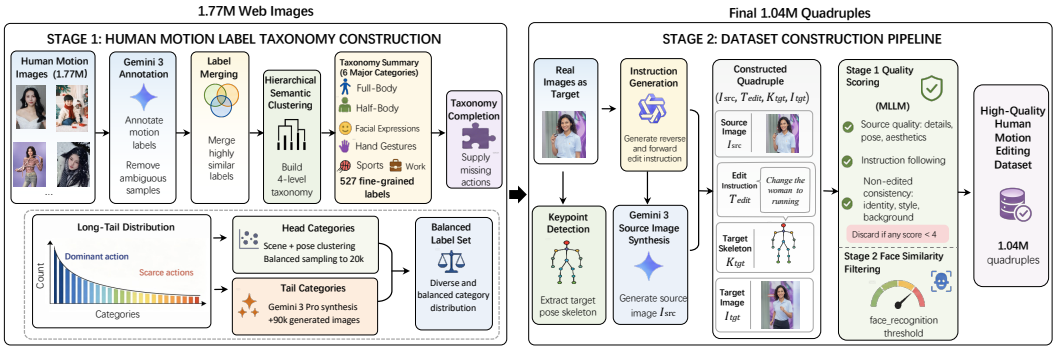

Figure 2: Pipeline of constructing the HME-1M dataset. We first build a hierarchical label taxonomy, which serves as the basis for balanced image sampling and supplementation, followed by the construction and quality control of image editing quadruplets.

Two-stage filtering pipeline. Motivated by the demonstrated efficacy of MLLM-based data assessment [32–34], the initial stage employs Qwen3-VL-8B [30] to score each sample pair along three critical dimensions using a five-point evaluation rubric. First, the overall quality of I_{src} is assessed based on generation details, pose coherence, and aesthetic composition. Second, for I_{tgt} as the edited result, we evaluate how well it follows the instruction. Third, we measure non-edited consistency, requiring strict preservation of identity-related attributes, as well as consistency in style and background (not applied to professional categories). Any pair receiving a score below 4 in any dimension is discarded. In the second stage, to maintain facial identity consistency, we compute face similarity between I_{src} and I_{tgt} using face_recognition [2], and remove pairs below a threshold of 0.97. Finally, we obtain 1.04M high-quality data pairs.

Human audit. To assess residual errors after filtering, we conducted a stratified human audit of 1,000 quadruplets, covering all six top-level categories and accounting for approximately 0.1% of HME-1M. Action-label errors, identity inconsistencies, and incorrect editing instructions were observed in 1.8%, 2.5%, and 6.6% of the audited samples, respectively.

3.3 HumanMotion Bench

After balanced sampling across scenes and pose characteristics, we select 10k real human images as evaluation candidates. To protect privacy, we replace faces using Gemini 3.0. We then retain images where the face similarity between the replaced and original faces falls between 0.88 and 0.96, as higher values compromise privacy and lower values produce unnatural averaged faces. Following the filtering procedure described in Section 3.2, we obtain a final subset of 1,581 images. Taking each taxonomy action as the target motion, we randomly select a source image from this subset and use Qwen3-VL-8B to generate three editing instructions per instance.

The substantial differences among motion categories pose challenges for accurately evaluating the motion quality of generated images. However, current benchmarks rely simply on "instruction following" to evaluate the quality of action execution , which fails to accurately measure motion completion. Consequently, we introduce three specialized, fine-grained metrics for each action category in section 5.2, ensuring that the assessment aligns closely with the critical quality dimensions specific to each action type.

4 Method

4.1 Overview

Human motion change is challenged by a semantickinematic [35] mismatch between high-level instructions and low-level joint dynamics. The single-step mapping paradigm becomes unreliable particularly when the target motion differs significantly from the original motion. Inspired by the

[2]https://github.com/ageitgey/face_recognition

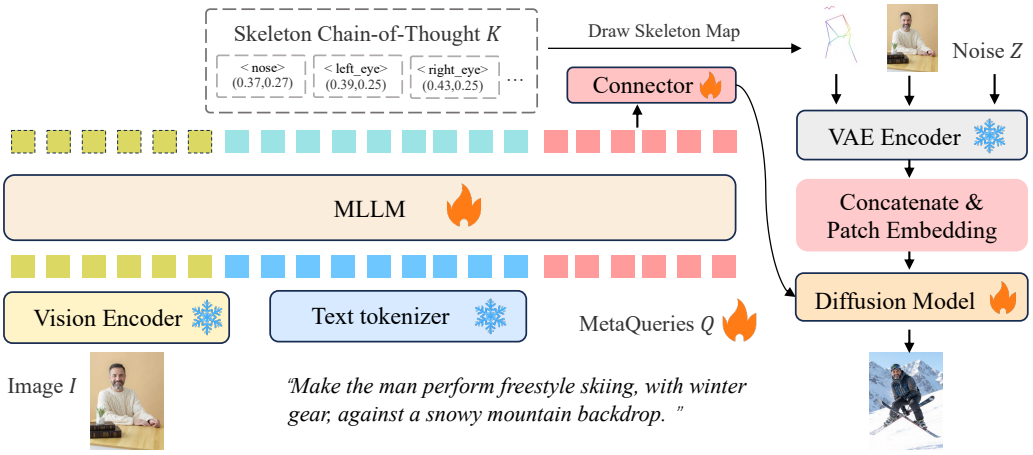

Figure 3: Illustration of our model architecture and the proposed Skeleton Chain-of-Thought (SCoT) framework. An MLLM acts as a pose planner to predict the target skeleton and MetaQuery features, which are provided as conditioning signals to the diffusion model to generate high-quality and strongly controllable images.

Think-Before-Generate paradigm [35–38], we restructure the human motion change task into a two-stage decision process: explicit topological planning followed by reference pose generation.

The overall architecture of our method is illustrated in Figure 3. In the first stage, the MLLM leverages chain-of-thought reasoning to predict the target skeleton coordinates as a motion prior. In the second stage, the skeleton map and the MetaQueries are jointly fed into the diffusion backbone. This architecture decouples intent planning from visual generation, effectively aligning instructions with physical kinematics and thereby enhancing instruction fidelity and environmental coherence.

4.2 Skeleton Chain-of-Thought

To construct a precise mapping from multimodal instructions to target motions, we utilize an MLLM $\mathcal{M}$ as an intermediate cognitive buffer. Specifically, it takes the source image I, the editing instruction T, and learnable query $\mathbf{Q}_{\text{init}} \in \mathbb{R}^{L\times d}$ as inputs, where L and d denote the number and dimension of the queries, respectively, to generate a heterogeneous planning sequence $\mathbf{y}$, formulated as:

$$P(\mathbf{y} \mid I, T) = \prod_{t=1}^{|\mathbf{y}|} P(y_t \mid y_{<t}, I, T), \tag{2}$$

where the sequence $\mathbf{y}$ is partitioned into a discrete coordinate sequence $\mathcal{K}$ and MetaQueries $\mathbf{Q} \in \mathbb{R}^{L\times d}$. Specifically, the sequence $\mathcal{K} = \{\langle k_j\rangle(x_j, y_j)\}_{j=1}^{N}$ adheres to the YOLO [39] keypoint annotation standard, where each element pairs a semantic joint identifier with normalized spatial coordinates (e.g., '<nose>(0.531, 0.226)'). We extract keypoint tuples from $\mathcal{K}$ using the regular expression parser $R(\cdot)$, which are then combined with the predefined skeletal topology connections T and rendered into a binary skeleton map, as

$$M = \text{Draw}(R(\mathcal{K}); T). \tag{3}$$

This process translates abstract instructions into concrete spatial geometric priors, effectively mitigating structural drift in limb configurations. Concurrently, $\mathbf{Q}$ encode deep reasoning features from the MLLM regarding global motion trends, high-level scene layout, and appearance consistency, thereby compensating for the contextual information.

4.3 Semantic and Structural Guidance for Generation

To enhance motion controllability and improve the visual fidelity of generated images, we propose to integrate implicit semantic guidance with fine-grained structural control for generation. Specifically, fine-grained information is encoded from I and M via a VAE encoder to obtain V_I and V_M,



Figure 4: Qualitative comparison of our method with state-of-the-art image editing models on Human Motion-Bench. Our method achieves the best performance in generating the target motion.

which are concatenated with the noise and fed as input to the DiT. The semantic control branch is implemented via MetaQuery features. In this branch, Q is passed through a trainable connector to obtain $Q' \in \mathbb{R}^{L\times c}$, which is then used as a conditioning input to the diffusion model D_θ.

The overall training objective is formulated as a weighted combination of the autoregressive loss $\mathcal{L}_{\text{AR}}$ for keypoint prediction and the flow-matching [40] loss $\mathcal{L}_{\text{FM}}$ for image generation, as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{AR}}\mathcal{L}_{\text{AR}} + \mathcal{L}_{\text{FM}}, \tag{4}$$

where λ_{AR} denotes the weight of the autoregressive loss.

In the early stages of training, inaccurate skeleton predictions would serve as erroneous conditions and negatively affect the training of D_θ. Therefore, during training we generate M using ground-truth skeletons, while at inference time we use that from prediction. Through end-to-end joint optimization, the planner $\mathcal{M}$ not only learns a precise mapping from the instruction to the target skeleton, but also fosters synergy between the implicit representations $\mathbf{Q}$ and the explicit coordinates $\mathcal{K}$. This ensures that the final generated images strictly adhere to the instructions while maintaining high physical plausibility and visual fidelity.

5 Experiments

5.1 Experimental Setup

We employ Ling-Mini-2.0 [41] as the MLLM backbone, which is a 16B-parameter model with 1.4B activated parameters, and adopt Z-Image [42] as the DiT. The number of MetaQueries is set to $L = 256$. λ_{AR} is set to 0.5. Using Fully Sharded Data Parallelism (FSDP) for distributed training, we train the model on the human motion dataset for one epoch on 8 NVIDIA H20.

In our experiments, we compare with 10 open-source state-of-the-art image editing models on HumanMotion Bench: Instruct-P2P [17], MagicBrush [43], AnyEdit [44], UltraEdit [45], Step1XEdit [6], BAGEL [9], UniWorld-V1 [46], FLUX.2-dev [8], MotionEdit [11] and Qwen-Image-Edit [7].

5.2 Metrics

Our evaluation protocol comprises three dimensions:



Figure 5: Our method supports user-defined skeleton inputs to enable customized human motion change. The skeleton map provides finer-grained control, including limb angles as well as full-body, half-body, and close-up pose control.

Generative Metrics. Following prior works [6, 11], we employ Gemini 3.0 to assess the generation quality of edited images, leveraging its advanced multimodal understanding capabilities and strong alignment with human preferences. The evaluation covers three aspects: (1) Identity Preservation, (2) Instruction Following (Fidelity), and (3) Overall Quality of the Edited Image (Quality).

Pairwise Preference. Inspired by GEdit-v2 [13], we adopt pairwise comparisons to better reflect human comparative preferences. Specifically, we use Gemini 3.0 to conduct head-to-head evaluations and report the win rate.

Fine-grained Metrics by Motion Category. Our fine-grained evaluation metrics (Motion) are defined as follows: (1) Full/Half-Body Editing, focusing on pose naturalness, limb motion fluency, and human-scene harmony; (2) Facial Expression Editing, assessing expression naturalness, emotional accuracy, and facial feature coherence; (3) Gesture Editing, evaluating anatomical hand correctness, gesture semantic clarity, and hand-body coordination; (4) Professional Sports Editing, measuring kinematic rationality, professional proficiency, and equipment-body interaction realism; and (5) Professional Work Editing, examining normative occupational actions, realistic tool interactions, and character-environment consistency. We define scoring criteria for each metric and use Gemini 3.0 to evaluate the generated images. The evaluation standards and prompts are provided in Appendix C.

5.3 Experimental Results

Qualitative Comparison Figure 4 compares our method with open-source state-of-the-art image editing models on HumanMotion Bench. For subtle and fine-grained editing instructions (rows 1, 3), our SCoT paradigm leverages the understanding capability of MLLMs to parse the instruction and generate the most appropriate motion as conditioning for the diffusion model, enabling stronger instruction-following in the generated motions. For editing instructions involving significant scene changes and more specialized target actions (rows 2, 4), our method produces professional-level motions while preserving identity consistency and achieving natural integration with the scene, yielding results comparable to the closed-source model Seedream5 [47].

By jointly leveraging explicit skeleton features and implicit MetaQuery features, our method allows users to additionally specify a reference skeleton during inference. Figure 5 illustrates the generation results of our approach. Other state-of-the-art methods, such as Qwen-Image [7] and MotionEdit [11], can only modify poses through natural language instructions, which often fails to satisfy users needs for fine-grained control. In contrast, the skeleton map provides more precise motion control: users can adjust limb angles by modifying the positions of keypoints (Rows 12), and control full-body, half-body, or close-up compositions through the visibility of the skeleton (Row 3).

Table 1: Performance comparison on the HumanMotion Bench.

Model	HumanMotion Bench					
	ID Preserv.	**Fidelity**	**Quality**	**Motion**	**Overall**	**Win Rate**
Instruct-P2P[17]	3.10	2.33	2.18	2.55	2.54	8.44%
AnyEdit [44]	4.74	1.95	2.58	3.73	3.25	12.26%
MagicBrush [43]	5.25	3.15	2.72	2.83	3.49	15.11%
UltraEdit [45]	1.26	2.85	2.33	2.83	2.32	7.43%
UniWorld-V1 [46]	5.59	4.04	6.02	6.07	5.43	26.08%
Step1X-Edit [6]	8.15	7.58	6.18	4.90	6.70	64.07%
BAGEL [9]	7.29	6.84	5.48	4.22	5.96	56.48%
FLUX.2-dev [8]	7.68	7.51	**7.17**	5.91	7.06	84.43%
MotionEdit [11]	4.55	6.61	3.89	3.60	4.66	51.33%
Qwen-Image [7]	7.14	7.63	7.08	5.72	6.89	73.24%
AnyMotion (Ours)	**8.92**	**7.87**	6.89	**6.42**	**7.52**	**85.09%**

Quantitative Comparison Table 1 reports the quantitative results of our method and other approaches on the HumanMotion Bench. Our method achieves the highest overall score, outperforming the second-best model FLUX.2-dev by 0.46. Benefiting from the strict control of identity-related attributes in HME-1M, our method achieves strong performance in identity preservation (**ID Preserv.**). In addition, by using SCoT to establish a precise mapping from user instructions to target actions, our method achieves stronger instruction-following capability (**Fidelity**). By incorporating explicit skeleton constraints, our method trades a slight but acceptable quality degradation for enhanced motion controllability. Consequently, its **Quality** score is only lower than those of Qwen-Image and FLUX.2-dev, while still surpassing all other methods. Furthermore, the large amount of high-quality real images used as supervision enables the model to generate more natural results when handling specific actions, which is reflected in higher **Motion** scores. Existing open-source SOTA models achieve relatively low scores on these fine-grained motion quality metrics, highlighting the value of HumanMotion Bench as a new benchmark that provides clear directions for future model improvements.

Alignment with human judgments. To validate our automated evaluation, we randomly sampled 2,000 pairwise comparisons across the six motion categories of HumanMotion Bench, stratified by motion-evaluation group. Five expert annotators independently compared outputs from AnyMotion and the evaluated baselines under a blinded protocol, using the same evaluation rubrics. Inter-annotator agreement was 91%, while agreement between Gemini and human pairwise judgments exceeded 87%. Human and Gemini evaluations also yielded the same overall ranking of the evaluated models. These results support the alignment of our automated evaluation with human preferences on the assessed subset.

Ablation Study To investigate the optimal strategy for incorporating skeleton keypoint features into DiT, we compare our AnyMotion with three variants. The "w/o SCOT" variant denotes the baseline configuration without keypoint coordinate prediction. The "w/o Skeleton Condition" variant predicts keypoints but lacks a dedicated motion control design. The "w PoseEncoder" variant employs a lightweight Transformer-based point cloud network to encode skeleton features. We evaluate these configurations in terms of both overall generation quality and average Relative Keypoint Deviation (RKD), and the quantitative results are summarized in Table 2.

Compared with generating MetaQueries alone as conditioning signals, predicting keypoints establishes an explicit mapping from instructions to concrete actions, improving the overall score by 0.49. Due to the inherent misalignment between point-cloud space and pixel space, PoseEncoder further improves performance by 0.77 but still struggles to achieve precise pose control. In contrast, our method enables more accurate skeleton guidance and achieves the best overall performance.

Table 2: Ablation study on different strategies for encoding skeleton points.

Method	ID Preserv.↑	Fidelity↑	Quality↑	Motion↑	Overall↑	RKD ↓
w/o SCOT	8.55	5.14	6.42	4.15	6.06	-
w/o Skeleton Concat	8.37	7.20	6.31	4.32	6.55	12.07%
w PoseEncoder	8.62	7.86	**6.93**	5.88	7.32	5.17%
AnyMotion	**8.92**	**7.87**	6.89	**6.42**	**7.52**	**1.43%**

6 Discussion

We introduce AnyMotion, a novel open-world paradigm that enables highly controllable, cross-scene motion synthesis. We construct HME-1M, a large-scale dataset comprising 1.04 million rigorously curated editing quadruples across 527 fine-grained motion categories, alongside a balanced sampling and quality control pipeline. Concurrently, we establish HumanMotion Bench, the first benchmark to provide category-specific, fine-grained evaluation protocols for human motion change. Furthermore, we propose the SCoT framework, which decouples high-level intent reasoning from pixel-level generation. By leveraging an MLLM as a cognitive planner to predict explicit target skeletons and implicit MetaQueries, SCoT effectively bridges the semantickinematic gap and supports precise, user-defined pose control. Experiments show that AnyMotion achieves accurate human motion editing across a variety of scenarios.

Limitations and Broader Impact Although our method achieves precise control over human motion, it is currently limited to the image level and cannot yet generate continuous motion edits at the video level. AnyMotion can significantly lower the barrier to human-centric content creation and provide a new benchmark for structured reasoning and controllable generation with multimodal large models. However, its high fidelity and strong controllability also introduce potential risks of misuse, as the technology could be exploited to produce deepfake content. Therefore, it should be accompanied by human oversight and fairness evaluation to ensure safe and controlled use.

References

[1] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2430–2449, 2023.

[2] Z. Zhang, Y. Wang, B. Wu, S. Chen, Z. Zhang, S. Huang, W. Zhang, M. Fang, L. Chen, and Y. Zhao, "Motion avatar: Generate human and animal avatars with arbitrary motion," *arXiv preprint arXiv:2405.11286*, 2024.

[3] R. Zhang, B. Yu, J. Min, Y. Xin, Z. Wei, J. N. Shi, M. Huang, X. Kong, N. L. Xin, S. Jiang *et al.*, "Generative ai for film creation: A survey of recent advances," *arXiv preprint arXiv:2504,08296*, 2025.

[4] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin *et al.*, "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[5] Z. Zhang, J. Liao, M. Li, Z. Dai, B. Qiu, S. Zhu, L. Qin, and W. Wang, "Tora: Trajectory-oriented diffusion transformer for video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 2063–2073.

[6] S. Liu, Y. Han, P. Xing, F. Yin, R. Wang, W. Cheng, J. Liao, Y. Wang, H. Fu, C. Han *et al.*, "Step1x-edit: A practical framework for general image editing," *arXiv preprint arXiv:2504.17761*, 2025.

[7] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.

[8] Black Forest Labs, "FLUX.2: Frontier Visual Intelligence," https://bfl.ai/blog/flux-2, 2025.

[9] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.

[10] N. Rotstein, G. Yona, D. Silver, R. Velich, D. Bensaïd, and R. Kimmel, "Pathways on the image manifold: Image editing via video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 7857–7866.

[11] Y. Wan, L. Ke, W. Yu, K.-W. Chang, and D. Yu, "Motionedit: Benchmarking and learning motion-centric image editing," *arXiv preprint arXiv:2512.10284*, 2025.

[12] M. Cao, X. Zhang, Y. Zheng, and Z. Xia, "Instruction-based image manipulation by watching how things move," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 2704–2713.

[13] Z. Jiang, Z. Sun, X. Zeng, Y. Yang, X. Zhang, Y. Wu, W. Cheng, G. Yu, X. Yang, and B. Wen, "Geditbench v2: A human-aligned benchmark for general image editing," *arXiv preprint arXiv:2603.28547*, 2026.

[14] X. Pan, S. N. Shukla, A. Singh, Z. Zhao, S. K. Mishra, J. Wang, Z. Xu, J. Chen, K. Li, F. Juefei-Xu *et al.*, "Transfer between modalities with metaqueries," *arXiv preprint arXiv:2504.06256*, 2025.

[15] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, S. Chen, and L. Cao, "Diffusion model-based image editing: A survey," *arXiv preprint arXiv:2402.17525*, 2024.

[16] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," in *International Conference on Learning Representations (ICLR)*, 2023.

[17] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 392–18 402.

[18] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[20] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[21] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "VideoComposer: Compositional video synthesis with motion controllability," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[22] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "GLIGEN: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[23] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 5, 2024, pp. 4296–4304.

[24] 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 (CVPR)*, 2022, pp. 10 684–10 695.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[26] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8162–8171.

[27] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 8780–8794.

[28] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4195–4205.

[29] Q. Gan, Y. Ren, C. Zhang, Z. Ye, P. Xie, X. Yin, Z. Yuan, B. Peng, and J. Zhu, "Humandit: Pose-guided diffusion transformer for long-form human motion video generation," *arXiv preprint arXiv:2502.04847*, 2025.

[30] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[31] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2024.

[32] E. Henriksson, O. Tarkka, and F. Ginter, "Finerweb-10bt: Refining web data with llm-based line-level filtering," in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, 2025, pp. 258–268.

[33] R. Chen, Y. Wu, L. Chen, G. Liu, Q. He, T. Xiong, C. Liu, J. Guo, and H. Huang, "Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 4156–4172.

[34] D. Cayir, R. Tao, R. Rungta, K. Sun, S. Chen, H. Khan, M. Kim, J. Reinspach, and Y. Liu, "Refine-n-judge: Curating high-quality preference chains for llm-fine-tuning," *arXiv preprint arXiv:2508.01543*, 2025.

[35] Y. Qian, J. Wang, Y. Feng, C. Xu, W. Lu, Y. Liu, B. Sun, Y. Chen, Y. Liu, and S. Wang, "Think before you move: Latent motion reasoning for text-to-motion generation," *arXiv preprint arXiv:2512.24100*, 2025.

[36] S. Kou, J. Jin, Z. Zhou, Y. Ma, Y. Wang, Q. Chen, P. Jiang, X. Yang, J. Zhu, K. Yu *et al.*, "Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders," *arXiv preprint arXiv:2601.10332*, 2026.

[37] S. Jiao, Y. Lin, Y. Zhong, Q. She, W. Zhou, X. Lan, Z. Huang, F. Yu, Y. Yu, Y. Zhao *et al.*, "Thinkgen: Generalized thinking for visual generation," *arXiv preprint arXiv:2512.23568*, 2025.

[38] C.-H. Yeh, Y. Wang, N. Zhao, R. Zhang, Y. Li, Y. Ma, and K. K. Singh, "Beyond simple edits: X-planner for complex instruction-based image editing," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 40, no. 14, 2026, pp. 11 991–11 999.

[39] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[40] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[41] A. Li, B. Liu, B. Hu, B. Li, B. Zeng, B. Ye, C. Tang, C. Tian, C. Huang, C. Zhang *et al.*, "Every activation boosted: Scaling general reasoner to 1 trillion open language foundation," *arXiv preprint aiXiv: 2510.22115*, 2025.

[42] H. Cai, S. Cao, R. Du, P. Gao, S. Hoi, Z. Hou, S. Huang, D. Jiang, X. Jin, L. Li *et al.*, "Z-image: An efficient image generation foundation model with single-stream diffusion transformer," *arXiv preprint arXiv:2511.22699*, 2025.

[43] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," vol. 36, pp. 31 428–31 449, 2023.

[44] Q. Yu, W. Chow, Z. Yue, K. Pan, Y. Wu, X. Wan, J. Li, S. Tang, H. Zhang, and Y. Zhuang, "Anyedit: Mastering unified high-quality image editing for any idea," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 26 125–26 135.

[45] H. Zhao, X. Ma, L. Chen, S. Si, R. Wu, K. An, P. Yu, M. Zhang, Q. Li, and B. Chang, "Ultraedit: Instruction-based fine-grained image editing at scale," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 3058–3093, 2024.

[46] B. Lin, Z. Li, X. Cheng, Y. Niu, Y. Ye, X. He, S. Yuan, W. Yu, S. Wang, Y. Ge *et al.*, "Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation," *arXiv preprint arXiv:2506.03147*, 2025.

[47] T. Seedream, Y. Chen, Y. Gao, L. Gong, M. Guo, Q. Guo, Z. Guo, X. Hou, W. Huang, Y. Huang *et al.*, "Seedream 4.0: Toward next-generation multimodal image generation," *arXiv preprint arXiv:2509.20427*, 2025.

A Additional Results



Figure 6: Qualitative comparison of our method with state-of-the-art image editing models on HumanMotion Bench.

In Figure 6, we present additional results of our method and other state-of-the-art image editing models on HumanMotion Bench. Our method achieves strong performance in instruction following, pose coordination, and target motion completion.

Figure 7 illustrates the fine-grained evaluation metrics we propose for different categories of human actions. Although existing state-of-the-art methods achieve promising results in facial and full-body editing, motion changes in professional work and sports scenarios remain insufficiently addressed. Therefore, our HumanMotion Bench provides a more targeted evaluation framework and highlights directions for future improvements.

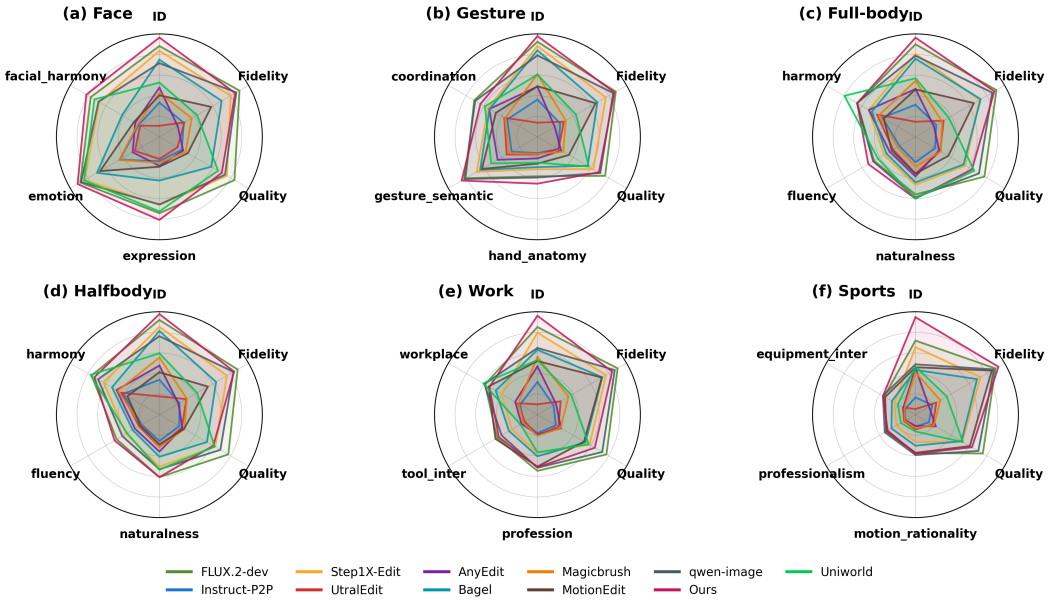

Figure 7: Detailed results of motion-specific metrics on HumanMotion Bench.

B Prompts for HME-1M Pipeline

Prompt for Generate Source Image

```
Generate an editing instruction based on the input Image A to guide another model
    to produce a new image B with a different human action.

The action of the person in Image A is: {action}
The editing instruction should meet the following requirements:
1. The instruction should focus on changing the persons action.
2. The target action in the instruction should be **significantly different**
    from the action in Image A. For example, if the person in Image A has an
    angry expression, the instruction can ask to change it to a smile; if the
    person in Image A is running, the instruction can ask to change it to
    walking.
3. The tone should sound natural, simulating how a user would give an editing
    request.

Please output only the instruction.
```

Prompt for Quality control

```
You are an expert in image editing quality evaluation.

Given Image 1 (the original image), an editing instruction, and Image 2 (the
    edited image), please evaluate the overall quality of this image pair.

The editing instruction is: {instruction}

Please rate the following aspects on a scale from 1 to 5 (5 is the highest and 1
    is the lowest):

1. quality_A1: How is the overall quality of Image A? (Consider body kinematics,
    aesthetic composition, and the proportional coordination of the human body.)
2. quality_A2: Is the human figure in Image A normal? (Focus on the number of
    hands and limbs, and whether the body follows physical and motion
    constraints.)
3. quality_AB: How consistent are the faces between Image A and Image B?
4. quality_B: As the edited result, how well does Image B follow the editing
    instruction?

Please output the scores in JSON format, with the keys: quality_A1, quality_A2,
    quality_AB, quality_B.
```

C Criterion and Prompts for HumanMotion Bench

Table 3: Evaluation metrics for portrait motion-change editing tasks. Universal metrics apply across all domains, while category-specific metrics address distinct editing requirements. Scores range from 0 to 10, with -1 indicating non-applicability.

Category	Metric	Concise Description
Universal	id consistency	Preserves identity-defining facial features, hairstyle, and markers for unambiguous post-edit recognition.
	instruction adherence	Faithfully executes specified motion/expression changes without introducing extraneous modifications.
	overall quality	Delivers sharp, artifact-free imagery with natural lighting and coherent background integration.
Full/Half-body	pose naturalness	Renders spontaneous, everyday postures devoid of artificial staging artifacts.
	motion fluency	Ensures biomechanically smooth transitions across sequential body joints.
	scene person harmony	Aligns subject pose with environmental semantics, lighting, and perspective geometry.
Facial Expression	expression naturalness	Generates anatomically plausible facial muscle dynamics with organic emotional transitions.
	emotion accuracy	Precisely conveys target affective states with appropriate intensity and contextual nuance.
	facial harmony	Coordinates synchronized dynamics across all facial regions for unified emotional expression.
Gesture Editing	hand anatomy correctness	Renders biomechanically plausible finger joints, proportions, and palm perspective.
	gesture semantic clarity	Produces culturally unambiguous hand signs aligned with user intent.
	hand body coordination	Integrates gesture dynamics with posture, gaze, and expression for behavioral coherence.
Professional Sports	motion dynamics rationality	Models sport-specific kinematics with accurate force chains and temporal sequencing.
	overall professionalism	Achieves elite-level athletic execution quality, power expression, and movement completion.
	equipment interaction realism	Simulates physically plausible gear-body contact, deformation, and light interaction.
Workplace Portrait	professional norm compliance	Depicts task execution conforming to occupational standards and industry conventions.
	tool interaction realism	Represents logically consistent manipulation and visual feedback with work instruments.
	workplace atmosphere consistency	Harmonizes subject engagement with office environment lighting and compositional mood.

Prompt for Pairwise Preference

```
You are an expert evaluator for motion image editing tasks.

**Edit Instruction**: "{safe_instruct}"

You will see an original image and TWO edited images (Edit A and Edit B).

Please compare Edit A and Edit B on the following criteria:
1. **Fidelity**: Which better executes the motion instruction?
2. **Preservation**: Which better preserves personal identity, background, and
    style?
3. **Coherence**: Which has more natural and physically plausible motion?

**Output Format** (JSON only):
{{
    "winner": "A" or "B" or "tie",
}}

Be objective and choose the better edit. If both are equally good/bad, choose
    "tie"."""
```

Prompt for Generative Metrics

```
You are a professional image quality evaluation expert. Please conduct a
    multi-dimensional quality assessment of the following portrait motion-change
    editing result.

[Input]
- Original portrait image
- Editing instruction: {INSTRUCTION}
- Edited/generated image

[Evaluation Dimensions and Scoring Criteria] (Each dimension is scored out of 10.
    Higher scores indicate better quality.)

1. Identity Preservation (ID Consistency)
- 10: Facial features, hairstyle, skin tone, and distinctive attributes (e.g.,
    moles, glasses, tattoos) are perfectly consistent; the identity is
    unambiguous.
- 7~9: Core identity features are well preserved; minor differences do not affect
    identity recognition.
- 4~6: Some features change, but the person is still recognizable as the same
    individual.
- 1~3: Facial structure or key features change significantly, making identity
    recognition difficult.
- 0: Completely unrecognizable as the same person.

2. Instruction Adherence
- 10: The motion instruction is executed precisely; pose, expression, and body
    changes fully match expectations with no unnecessary modifications.
- 7~9: The main action requirements are satisfied; minor deviations in details do
    not affect the overall result.
- 4~6: The instruction is only partially followed; the motion shows noticeable
    deviations or misses key elements.
- 1~3: Only weakly reflects the intent of the instruction; the main motion does
    not match the requirement.
- 0: The editing instruction is completely ignored.

3. Overall Image Quality of the Edited Result
- 10: The image is clear and sharp, lighting is natural, and there are no
    artifacts, distortions, or blur; the background and subject are well
    integrated.
- 7~9: Overall visual quality is good, with minor imperfections that do not
    affect usability.
- 4~6: Noticeable noise, edge artifacts, or inconsistent lighting exist, but the
    main subject remains acceptable.
- 1~3: Poor image quality with severe artifacts, blur, or distortion affecting
    visual perception.
- 0: The image is severely corrupted and unusable.

[Scoring Requirements]
- Please assign independent scores for each dimension based on professional
    visual judgment.
- Scores range from 0 to 10 and may include one decimal place (e.g., 8.5). If a
    dimension is not applicable, output -1.
- Output only the result in JSON format, without any explanation, analysis, or
    additional text.

[Output Format] (Strictly follow this format and return only the JSON)

{
  "id_consistency": <score>,
  "instruction_adherence": <score>,
  "overall_quality": <score>
}
```

Prompt for Evaluation of Facial Expression Editing

```
You are an expert in facial expression editing quality evaluation. Please
    evaluate the generated image only for facial expression editing, using the
    following three fine-grained metrics.

[Input]

Original portrait image
Editing instruction: {INSTRUCTION}
Edited/generated image
[Evaluation Metrics] (Each metric ranges from 110. If the metric is not relevant
    to the current editing task, output -1.)

expression_naturalness (Expression Naturalness)
10: Facial muscle movements fully follow anatomical principles; the expression
    transition is natural with no stiffness.
7~9: Overall natural; only minor stiffness in small details that does not affect
    visual perception.
4~6: Noticeable stiffness or distortion exists, but the main expression is still
    recognizable.
1~3: The expression is severely distorted; facial muscle movements violate
    physiological principles.
-1: The current editing task does not involve expression changes (e.g., only pose
    editing).
emotion_accuracy (Emotion Expression Accuracy)
10: The generated expression perfectly matches the target emotion (e.g.,
    happiness, anger, sadness, surprise); composite emotions are expressed
    precisely.
7~9: The primary emotion is correct, with minor deviations in secondary emotional
    details.
4~6: The emotional direction is correct, but the intensity or mixture is
    inaccurate.
1~3: The emotional expression significantly deviates from the instruction or
    becomes ambiguous.
-1: The instruction does not specify a target emotion.
facial_harmony (Facial Feature Harmony)
10: The dynamic coordination among eyes, eyebrows, nose, mouth, and cheeks is
    natural and consistent with real emotional expression.
7~9: Facial features are well coordinated, with only one minor inconsistency.
4~6: Some facial features change asynchronously, with local over-adjustment.
1~3: Facial dynamics are fragmented, producing a "collage-like" or contradictory
    expression.
-1: The editing task does not involve changes in facial feature dynamics.
[Output Requirements]

Output only JSON, without any explanation, analysis, markdown, or extra text.
Strictly use the following key names.
[Output Format]
{
"expression_naturalness": ,
"emotion_accuracy": ,
"facial_harmony":
}
```

Prompt for Evaluation of Body Editing

```
You are an expert in everyday portrait editing quality evaluation. Please
     evaluate the generated image **only for full-body or half-body everyday
     photo editing tasks**, using the following three fine-grained metrics.

[Evaluation Metrics] (Each metric ranges from 110. If the metric is not relevant
     to the current editing task, output -1.)

1. pose_naturalness (Naturalness of Everyday Pose)
- 10: The pose looks like a genuine candid capture, fully consistent with natural
     daily behavior and showing no signs of staged posing.
- 7~9: Overall natural; slight signs of intentional posing may exist but do not
     affect the everyday feeling.
- 4~6: Noticeable stiffness or staged appearance exists, but the pose itself
     remains reasonable.
- 1~3: The action appears exaggerated or distorted and does not resemble a
     realistic everyday scenario.
- -1: The editing task does not involve pose or action changes.

2. motion_fluency (Motion Fluency of Limbs)
- 10: Transitions across headshoulderhipleg joints are smooth and coherent; the
     motion has a complete and logical flow with no discontinuities.
- 7~9: Generally fluent; only one joint transition appears slightly stiff.
- 4~6: Noticeable joint jumps or pose discontinuities exist, but the overall
     motion remains coherent.
- 1~3: Limb connections appear twisted or broken, and the motion logic is chaotic.
- -1: The editing task does not involve multi-joint coordinated motion.

3. scene_person_harmony (ScenePerson Harmony)
- 10: The persons motion is fully consistent with the background in terms of
     scene semantics, lighting, and perspective.
- 7~9: Good integration overall; minor inconsistencies in lighting or scale may
     exist.
- 4~6: Slight separation between the person and the scene is noticeable but still
     acceptable.
- 1~3: The person appears pasted onto the background, with severely distorted
     spatial relationships.
- -1: The editing task does not involve background or scene elements.

[Output Requirements]
- Output **only JSON**, without any explanation, analysis, markdown, or
     additional text.
- Scores can be integers or one decimal place, within the range: -1, 1, 2, ...,
     10.
- Strictly use the following key names.

[Output Format]
{
  "pose_naturalness": <score>,
  "motion_fluency": <score>,
  "scene_person_harmony": <score>
}
```

Prompt for Evaluation of Gesture Editing

```
You are an expert in gesture editing quality evaluation. Please evaluate the
    generated image **only for gesture editing tasks**, using the following
    three fine-grained metrics.

[Evaluation Metrics] (Each metric ranges from 110. If the metric is not relevant
    to the current editing task, output -1.)

1. hand_anatomy_correctness (Correctness of Hand Anatomy)
- 10: Finger joint angles, finger segment proportions, and palm perspective fully
    conform to human ergonomics.
- 7~9: The structure is mostly correct; minor proportional deviations do not
    affect recognition.
- 4~6: One or two joints show distortion or proportional errors, but the main
    gesture remains identifiable.
- 1~3: Severe structural errors in the hand, such as extra fingers, reversed
    joints, fused fingers, or other artifacts.
- -1: The editing task does not involve hand or gesture changes.

2. gesture_semantic_clarity (Semantic Clarity of the Gesture)
- 10: The gesture meaning is clear and unambiguous, consistent with the intended
    cultural context and the instruction.
- 7~9: The main meaning is clear; minor ambiguities exist but do not affect
    understanding.
- 4~6: The gesture is somewhat ambiguous and requires contextual interpretation.
- 1~3: The gesture meaning contradicts the instruction or leads to cultural
    misinterpretation.
- -1: The instruction does not specify a specific gesture semantic target.

3. hand_body_coordination (HandBody Coordination)
- 10: The gesture is dynamically consistent with body posture, gaze, and facial
    expression, forming a natural behavioral chain.
- 7~9: Good coordination overall; only minor linkage deviations.
- 4~6: Slight disconnection between the gesture and other body parts.
- 1~3: The gesture appears isolated or abrupt and conflicts with the overall
    posture.
- -1: The editing task does not involve coordination with full-body posture.

[Output Requirements]
- Output **only JSON**, without any explanation, analysis, markdown, or
    additional text.
- Scores can be integers or one decimal place, within the range: -1, 1, 2, ...,
    10.
- Strictly use the following key names.

[Output Format]
{
  "hand_anatomy_correctness": <score>,
  "gesture_semantic_clarity": <score>,
  "hand_body_coordination": <score>
}
```

Prompt for Evaluation of Sports

```
You are a professional sports motion editing quality evaluation expert. Please
    evaluate the generated image **only for professional sports editing tasks**,
    using the following three fine-grained metrics.

[Evaluation Metrics] (Each metric ranges from 110. If the metric is not relevant
    to the current editing task, output -1.)

1. motion_dynamics_rationality (Rationality of Motion Dynamics)
- 10: The motion fully conforms to the technical standards of the sport; the
    force chain, center of mass, and timing are precise.
- 7~9: The main movement is reasonable, with only minor deviations in secondary
    details.
- 4~6: Noticeable technical errors exist, but the type of sport motion remains
    recognizable.
- 1~3: The motion violates the principles of sports biomechanics, showing
    unnatural or "impossible" postures.
- -1: The editing task does not involve professional sports actions.

2. overall_professionalism (Overall Professionalism of the Motion)
- 10: The rhythm, power, and completeness of the motion reach a professional
    athlete level, with strong dynamic tension.
- 7~9: The motion demonstrates good professionalism, with only minor deviations
    from competitive standards.
- 4~6: The motion has recognizable sports characteristics but appears amateurish
    and lacks professional details.
- 1~3: The motion is loose and weak, lacking the characteristics of professional
    sports performance.
- -1: The instruction does not require demonstrating a professional sports level.

3. equipment_interaction_realism (Realism of EquipmentBody Interaction)
- 10: The contact relationships between equipment/clothing and the body,
    deformation feedback, and lighting/shadow interactions are fully physically
    consistent.
- 7~9: The interaction is mostly realistic, with only minor issues in deformation
    or lighting.
- 4~6: Noticeable clipping, floating, or incorrect deformation direction exists,
    but the overall result remains acceptable.
- 1~3: The relationship between equipment and body is severely distorted,
    breaking the realism of the motion.
- -1: The editing task does not involve interaction with sports equipment.

[Output Requirements]
- Output **only JSON**, without any explanation, analysis, markdown, or
    additional text.
- Scores can be integers or one decimal place, within the range: -1, 1, 2, ...,
    10.
- Strictly use the following key names.

[Output Format]
{
  "motion_dynamics_rationality": <score>,
  "overall_professionalism": <score>,
  "equipment_interaction_realism": <score>
}
```

Prompt for Evaluation of Work Actions

```
You are an expert in workplace portrait editing quality evaluation. Please
    evaluate the generated image **only for professional work editing tasks**,
    using the following three fine-grained metrics.

[Evaluation Metrics] (Each metric ranges from 110. If the metric is not relevant
    to the current editing task, output -1.)

1. professional_norm_compliance (Compliance with Professional Action Norms)
- 10: The posture and operations fully follow the standard operating procedures
    (SOP) and industry conventions of the profession; professionalism is
    excellent.
- 7~9: The main actions follow professional norms, with only minor deviations
    from standard procedures.
- 4~6: Noticeable professional action errors exist, but the occupational role is
    still recognizable.
- 1~3: The actions seriously violate professional norms, creating an
    unprofessional or awkward impression.
- -1: The editing task does not involve profession-specific actions.

2. tool_interaction_realism (Realism of Work Tool Interaction)
- 10: Interactions with computers, tools, documents, or other work items are
    fully realistic in terms of logic, contact relationships, and lighting
    feedback.
- 7~9: The interaction is mostly realistic, with only minor flaws in operational
    details.
- 4~6: Visible clipping, floating objects, or incorrect operational logic exist,
    but the main interaction remains recognizable.
- 1~3: The relationship between tools and the person is severely distorted,
    undermining the credibility of the work scene.
- -1: The editing task does not involve interaction with work tools.

3. workplace_atmosphere_consistency (Consistency of Workplace Atmosphere)
- 10: The persons state (e.g., focusing, communicating, presenting) is highly
    consistent with the office environment in lighting, composition, and color
    tone.
- 7~9: The atmosphere is well coordinated, with only slight deviations in color
    tone or depth-of-field.
- 4~6: A mild separation exists between the persons state and the scene, but the
    overall workplace feeling remains acceptable.
- 1~3: The persons emotion or posture strongly conflicts with the office
    environment, resulting in a chaotic atmosphere.
- -1: The editing task does not involve workplace environment or atmosphere.

[Output Requirements]
- Output **only JSON**, without any explanation, analysis, markdown, or
    additional text.
- Scores can be integers or one decimal place, within the range: -1, 1, 2, ...,
    10.
- Strictly use the following key names.

[Output Format]
{
  "professional_norm_compliance": <score>,
  "tool_interaction_realism": <score>,
  "workplace_atmosphere_consistency": <score>
}
```

D Label Taxonomy

In this section, we present the third-level labels and all corresponding fourth-level labels in our dataset.

Full Body - Interactive Actions: hug, support, assist, lift up, carry on the back, bow, play, snuggle, take a photo together, kiss, talk.

Full Body - Locomotor Movements: walk, catwalk, run, run backward, crawl, jump, jog, stride, tiptoe, pace, hop, go up stairs, go down stairs, climb.

Full Body - Pose: lean, sit down, stand against wall, stand with arms crossed, stand with hands on hips, stand, lie prone, sit with knees hugged, squat, lie down, sit cross-legged, put hands on head, kneel, sleep, sit with legs outstretched, bend over, stretch, stand on one leg, squat with knees hugged, lie on side, crawl low, lie on back, stand on tiptoe, lunge, fall down, dodge, horse stance, pose, put hands in pockets, cross hands, hold up, put hands behind back, hold up with one hand, look back, handstand, balance object on head, raise hand, lift skirt, hold foot, carry object on head, observe, show muscle, put hand in pocket and hold object, step on, drape, kick leg, swing, turn around, hold flower in mouth and look at camera, wrap in blanket.

Half Body - Interactive Actions: place both hands under chin, touch hair, touch chest, touch another person's arm, touch back of neck, wave, place hand on table, make inviting gesture, cover mouth and nose, hug, place hand on another person's arm, push another person's shoulder, receive object, cross arms in front of chest, press palms together, place hand on knee, make stop gesture, make shushing gesture, place hand on chair back, adjust glasses, hold another person's wrist, pull hem, pull sleeve, touch face, place arm on shoulder, touch head, touch lips, touch chin, touch ear, touch tongue, touch belly, place hand on head, touch arm, tie a tie, touch body, listen, observe, reach out, raise both hands and touch, place one hand in front of body, place hand, place right hand on collar, swing hand, pinch cheek, place hand on face and tilt head, adjust clothes, touch eyes, place hand on branch, touch forehead, hand interaction, touch ear, touch nose, wipe sweat, drink water.

Half Body - Posture Adjustment: turn head, lean forward, lower head, raise head, lean sideways, shrug shoulders, hunch back, lean backward, sit upright, stretch neck, raise hand, cross hands in front of body, put hands in pocket, put hands behind back, extend arm

Half Body - Emotional Expression: cross arms in front of chest (defensive), pat forehead (annoyed), spread both hands (helpless), clasp hands tightly (anxious), cover face (ashamed), put hands on hips (angry), put hands on hips (confident), cover face and cry, bite finger (nervous), cover ears with both hands (refuse to listen), twirl hair with finger (thinking), rub hands repeatedly (nervous), frown and stare (focused), frown and stare (displeased), raise chin slightly (arrogant), lower head and stay silent (ashamed), put hands on head (frustrated), hold fists in front of chest (prayer-like anxiety), hug oneself (uneasy), sigh with dropping shoulders, bite lip (hesitant), turn mouth corners down (disappointed), tap fingers on table (impatient), grab hair (irritated), pound chest (regretful), pound chest (emphatic), shake head and sigh, avoid eye contact (lower head), avoid eye contact (turn head),close eyes tightly (resistant).

Gesture - Fine Hand Movements: apply ointment, pinch, pluck strings, play the piano, play the flute, play the trumpet, paint, use tweezers, write with a Chinese brush, trim nails, type, fold paper, tighten screw, arrange flowers, use mouse, comb hair, thread needle, button clothes, weave, turn pages, untie shoelace, carve, groom, play the violin, play the guitar, play with fingers, make craft play the bass, play the cello, brush teeth, medical care, cut hair, snap fingers, bandage, apply face mask.

Gesture - Communicative Gestures: make peace sign, cover face, press palms together, make heart gesture, cover mouth, point at person, give thumbs up, place index finger on lips, make fist, cup hands in salute, make OK sign, make camera gesture, raise middle finger, make phone gesture, spread both hands, raise hand to ask question, raise both hands and wave, make writing gesture, raise hands in surrender, bow with hands clasped, wave goodbye, make victory sign, make X sign, pray, point, raise hand to signal, make hand gesture, salute, make claw gesture, make gesture near cheek, raise index finger, clap, blow kiss, touch index fingers together, raise glass, form hand seal, make finger gun, make rock sign, make glasses gesture, reach out, make triangle gesture, make bunny ears, make L sign, give thumbs down, raise hand, make peacock gesture, make orchid finger gesture.

Gesture - Operational Gestures: hold object, show object, hold umbrella, read book, write, draw, adjust watch, adjust headphones, pick up food, hold cup, spin prayer wheel, pour liquid, look at phone, make phone call, stir, grab, smoke, press, look in mirror, hold steering wheel, lift up, wear helmet, rub, use magnifying glass, clench fist, shave, tap, play chess or cards, turn palm over, take a selfie, adjust clothes, eat, drink, take photo, hold hat, drink coke, arrange object, lift skirt hem, play, drink wine, use electronic device, hold object in arms, raise glass, clink glasses, cook, open object, take pose, clean, open door, play with toy, play game, use weapon, splash water, spray, look at watch, handle object, wear accessory, adjust accessory, lift curtain, lift veil, adjust mask, touch face with object, set off confetti, throw, drink sprite, brush teeth, listen to music, wear glasses, scatter object, light cigarette, wear gloves.

Face - Mouth Motions: stick out tongue, chew, blow, bite, bite lip, yawn, lick lips, roll tongue, make tut sound, suck, drink through straw, blow bubble, chew gum, mumble, open mouth and breathe deeply, tear package with teeth, hold object in mouth, pout, speak, open mouth, lick, shout, press lips together, sing, curl lip.

Face - Head Motions: tilt head left, lower head, raise head, shake head, nod, lean head forward.

Face - Eye Motions: stare, blink, wink, squint sideways, open eyes wide, half-open eyes (sleepy), close eyes tightly, raise eyebrows, narrow eyes, move eyes up and down, move eyes side to side, focus eyes on near object, focus eyes on distant object, show nystagmus (small eye tremor), twitch eyelid, roll eyes, close eyes.

Face - Facial Expressions: smile, expressionless, pout angrily, laugh, open mouth in surprise, cry, curl lip in contempt, smile with narrowed eyes, wrinkle nose in disgust, grin, clench teeth, shrink face in fear, glare angrily with tense face, sob, laugh wildly, lower head and smile shyly, puff cheeks angrily, make face, frown, act cute, look thoughtful, enjoy and wink, frown and close eyes, slight frown, look aggrieved, look sad, look shocked, look confused, bitter smile, tearful stare.

Sports - Ball Sports: golf, football, ping pong, bowling, ice hockey, billiards, softball, squash, handball, volleyball, field hockey, cricket, baseball, rugby, water polo, beach volleyball, basketball, tennis, badminton, lacrosse, frisbee, darts, shot put, hammer throw, hold racket, swing racket, hold ball, throw ball, catch ball.

Sports - Athletics: run, sprint, middle-distance running, long-distance running, orienteering, tug of war, javelin, race walking, cross-country skiing, hurdles, jump rope, long jump, high jump, triathlon, shot put, marathon.

Sports - Equipment Sports: cycling, fencing, martial arts, fitness, dance, archery, yoga, weightlifting, equestrian, skateboarding, mountaineering, shooting, racing, wrestling, boxing, gymnastics, kendo, spinning, cheerleading, motocross, rock climbing, sanda, pilates, skydiving, judo, paragliding, esports, karate, trampoline, fishing.

Sports - Water Sports: swimming, rowing, surfing, sailing, paddleboarding, water skiing, diving, kayaking, dragon boat, water aerobics, aquatics-diving.

Sports - Winter Sports: skiing, biathlon, curling, snowboarding, freestyle skiing, figure skating, ski jumping, alpine skiing.

Work - Interpersonal Interaction: meeting discussion, customer consultation, first aid, guidance, directing, teaching, negotiation, greeting guests, bowing, interview, being interviewed, inspection.

Work - Device Interaction: drive, dig, warehouse inventory check, logging, standing guard, field sowing, parcel sorting, carrying, operating CNC machine, using microscope, circuit welding, breaking wall, programming, sewing, conducting experiment, gardening.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have discussed the limitations in the last Section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed explanations in the "experimental setup" and 'method' sections.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: In response to reviewers' concerns about potential privacy leakage and model misuse, we will not publicly release the HME-1M training dataset. We will release only HumanMotion Bench to support standardized evaluation while reducing the privacy and misuse risks associated with distributing large-scale human image data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided detailed explanations in the sections "Experiment" and "Method".

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not involve statistical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided an explanation in the 'Experimental Setup' section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We conducted a discussion in the "Discussion" section of the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: Due to privacy restrictions and to prevent potential risks of abuse, we have not open-source the training data.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.