VLM-Guided Group Preference Alignment for Diffusion-based Human Mesh Recovery

Supplementary Material

A. Details of the Primary Derivation

We derive the complete objective used in our method, starting from the original GRPO formulation, including the KL regularizer that was omitted in the main paper. For each condition $\boldsymbol{c}$, GRPO [17] draws a group of G samples $\mathcal{G} = \{\mathbf{x}_0^1, \dots, \mathbf{x}_0^G\}$ from the old policy. The GRPO update maximizes advantage-weighted likelihood ratios while keeping the updated policy close to a frozen reference policy p_{ref}, leading to the following objective (with clipping operation omitted):

$$-\mathbb{E}_{\boldsymbol{c},\,\mathcal{G}}\left[\sum_{i=1}^{G}\log\frac{p_\theta(\mathbf{x}_0^i|\boldsymbol{c})}{p_{\text{ref}}(\mathbf{x}_0^i|\boldsymbol{c})}A(\mathbf{x}_0^i)\right] \;+\; \beta_{\text{KL}}\,\text{KL}(p_\theta \,\|\, p_{\text{ref}})\,. \tag{1}$$

We construct the group-wise HMR preference dataset $\mathcal{G}_{\text{HMR}} = \big\{\big(I, (\mathbf{m}^1, s^1), (\mathbf{m}^2, s^2), \dots, (\mathbf{m}^G, s^G)\big)\big\}$ using the denoising outputs of a diffusion-based HMR base model, where each mesh prediction $\mathbf{m}^i$ is treated as the endpoint of a reverse diffusion trajectory $\mathbf{x}_{0:T}^i$. The log-likelihood ratio in Eq. (1) is intractable for diffusion models, so we follow the standard relaxation used in Diffusion-DPO [20] and approximate

$$\log\frac{p_\theta(\mathbf{m}^i|\boldsymbol{c})}{p_{\text{ref}}(\mathbf{m}^i|\boldsymbol{c})} \approx \mathbb{E}_{q(\mathbf{x}_{1:T}^i|\mathbf{m}^i)}\log\frac{p_\theta(\mathbf{x}_{0:T}^i|\boldsymbol{c})}{p_{\text{ref}}(\mathbf{x}_{0:T}^i|\boldsymbol{c})}, \tag{2}$$

where the forward process q is used as an approximation to the posterior over diffusion trajectories. Under the Gaussian parameterization of the reverse Markov chain, both p_θ and p_{ref} are uniquely defined by their noise predictors $\boldsymbol{\epsilon}_\theta$ and $\boldsymbol{\epsilon}_{\text{ref}}$. Following [8, 20], the path-space log-ratio admits the surrogate (with condition $\boldsymbol{c}$ omitted hereinafter)

$$\log\frac{p_\theta(\mathbf{m}^i|\boldsymbol{c})}{p_{\text{ref}}(\mathbf{m}^i|\boldsymbol{c})} \approx T\lambda_t\,\mathbb{E}_{t,\boldsymbol{\epsilon}}\big[L_{\text{DM}}^{\text{ref}}(\mathbf{x}_t^i,\boldsymbol{\epsilon}) - L_{\text{DM}}^{\theta}(\mathbf{x}_t^i,\boldsymbol{\epsilon})\big], \tag{3}$$

where $t \sim \mathcal{U}(1,T)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})$, and

$$L_{\text{DM}}^{\theta}(\mathbf{x}_t^i,\boldsymbol{\epsilon}) = \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^i,t) - \boldsymbol{\epsilon}\|_2^2. \tag{4}$$

The KL regularizer in Eq. (1) also admits a diffusion-space approximation. Under the same Gaussian reverse-kernel parameterization used in Diffusion-DPO [20], the trajectory-level KL between two diffusion policies admits a standard surrogate in noise-prediction space, yielding a weighted L_2 penalty between the noise predictors:

$$\text{KL}(p_\theta\|p_{\text{ref}}) \approx T\kappa_t\,\mathbb{E}_{t,\mathbf{x}_t}\big\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t) - \boldsymbol{\epsilon}_{\text{ref}}(\mathbf{x}_t,t)\big\|_2^2. \tag{5}$$

Here κ_t is a timestep-dependent weight for the KL term.

Substituting the diffusion surrogate Eq. (3) and the KL approximation Eq. (5) into the GRPO objective Eq. (1) yields our final training loss:

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{m}\sim\mathcal{G}_{\text{HMR}},\,t\sim\mathcal{U}(1,T),\,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\\ T\sum_{i=1}^{G}\Big[\beta\lambda_t A(\mathbf{m}^i)\left(L_{\text{DM}}^{\theta}(\mathbf{x}_t^i,\boldsymbol{\epsilon}) - L_{\text{DM}}^{\text{ref}}(\mathbf{x}_t^i,\boldsymbol{\epsilon})\right)\\ +\,\kappa_t\left\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^i,t) - \boldsymbol{\epsilon}_{\text{ref}}(\mathbf{x}_t^i,t)\right\|_2^2\Big],\end{aligned} \tag{6}$$

Although Eq. (6) includes the KL-derived surrogate term, in practice, we set its weight $\kappa_t = 0$ and optimize only the advantage-weighted diffusion loss in Eq. (**??**). Given that initialization from the reference model together with small-step fine-tuning already provides sufficient implicit regularization, and explicit KL did not bring consistent gains, we omit it in our implementation and main exposition.

B. Datasets

InstaVariety [10] is a 2D in-the-wild dataset comprising 2.1 million images gathered from Instagram using 84 distinct hashtags. The 2D annotation is generated from OpenPose [5].

Human3.6M [9] is a large-scale indoor motion-capture dataset containing 3.6M video frames with accurate 3D pose and shape annotations. Following standard practice [12, 21], we use subjects S1, S5, S6, S7, and S8 for training, and evaluate on subjects S9 and S11.

3DPW [19] is one of the most widely used in-the-wild benchmarks, which is captured with a moving phone and IMU sensors, providing accurate SMPL annotations across 60 video sequences recorded in diverse environments. We use the official IMU-derived SMPL parameters. Following [7, 18, 21], we train on the official 3DPW training split and evaluate on its official test split.

MPI-INF-3DHP [15] is a markerless motion-capture dataset covering both indoor and outdoor scenes. It records 8 subjects performing 8 activities from 14 camera views and provides accurate 3D joint annotations for evaluating cross-scene generalization.

MPII [1] is a widely used in-the-wild dataset containing around 25K images and over 40K annotated person instances. Approximately 28K instances are used for training, and the remaining samples form the test set.

COCO [14] is a large-scale dataset for object detection, segmentation, captioning, and 2D human keypoint detec-

tion. Its keypoint subset includes over 200K images and more than 250K annotated person instances.

UP-3D [13] is an in-the-wild dataset with 7,126 images. It provides high-quality 3D pseudo-annotations by extending SMPLify [3] to fit SMPL parameters for each image.

HI4D [22] is a human–human interaction dataset with rich physical contact. It provides 4D textured scans, SMPL annotations, vertex-level contact labels, and segmentation masks. The dataset contains 20 subject pairs over 100 sequences, totaling more than 11K frames.

BEDLAM [2] is a large-scale synthetic dataset with diverse body shapes, motions, skin tones, hair, and clothing. It is generated using 271 body models, 27 hairstyles, and 111 clothing types, rendered across various scenes with multiple people and camera viewpoints.

DNA-Rendering [6] is a large-scale, high-resolution multi-view studio dataset with diverse motions, clothing, and object interactions. It contains over 1.5K human instances and 5K motion sequences, captured from up to 60 RGB and 4 Kinect views, with annotations obtained via SMPLify-X [16] and multi-view optimization. For the evaluation of the critique model, we randomly sample 2,079 instances from the test set.

GTA-Human II [4] is a synthetic dataset built in GTA-V. It expands the original single-person GTA-Human dataset to multi-human scenes, providing 1.8M SMPL-X annotations generated with SMPLify-X [16] via MMHuman3D. Images are rendered in 4K with around 600 subjects performing diverse motions across varied backgrounds and viewpoints. For the evaluation of the critique model, we randomly sample 1,988 instances from the test split.

SPEC [11] is a synthetic dataset with diverse and unique camera viewpoints. It provides 22K images with 72K SMPL-labeled instances for training and 3.8K images with 12K instances for testing.

C. Prompts of Critique Agent

In Tables 1 and 2, we present the primary prompts used by our critique agent.

D. Limitations

Our method demonstrates strong generalization across extreme poses, heavy occlusions, complex backgrounds, and diverse unseen internet images. However, extreme camera viewpoints remain challenging, likely due to limited coverage of such cases in the training data, as illustrated in Figure 1. We expect that incorporating more diverse viewpoint data would further mitigate this limitation.

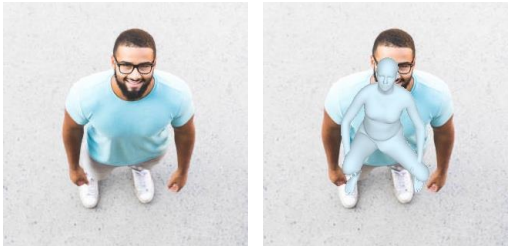

Figure 1. Failure case under extreme camera angles. This mainly arises from the lack of such viewpoints in the training data.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[4] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[6] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023. 2

[7] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21148–21158, 2023. 1

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1

[10] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5614–5623, 2019. 1

[11] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 2

[12] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 1

[13] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[15] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 1

[16] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[17] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[18] Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization. *arXiv preprint arXiv:2505.10250*, 2025. 1

[19] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1

[20] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 1

[21] Yuan Xu, Xiaoxuan Ma, Jiajun Su, Wentao Zhu, Yu Qiao, and Yizhou Wang. Scorehypo: Probabilistic human mesh estimation with hypothesis scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–989, 2024. 1

[22] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 2

Prompt Template: Critique Agent Scoring

```
You are an expert in 3D human mesh reconstruction and pose estimation.
The first image is an original image of a real person performing an action, with no 3D mesh
prediction overlaid.  Each subsequent image shows a predicted 3D human mesh overlaid on the
original image.
Please evaluate the prediction quality of each predicted image from 0 to 100, where the goal is
to clearly separate the quality between images | even small visual differences should lead to
noticeably different scores.
Do not score the original image.  Only provide scores for the predicted images.

For each predicted image:
• Give a score (0-100).
• Write a short comment describing the main issue or strength.
• Ensure the differences between scores are clearly noticeable.
Return the result strictly in JSON format like this:
{ "image_2":  {"score":  91, "comment":  "Accurate torso and limb alignment; only slight hand
drift."},
"image_3":  {"score":  77, "comment":  "Pose plausible but legs slightly floating off the
surface."},
"image_4":  {"score":  30, "comment":  "Wrong body orientation and poor foot contact."} }

Output only the JSON, nothing else.  <Input Images & Predicted Overlays>

You are a geometry-aware VLM judge for 3D human mesh predictions.
Read the following MEMORY CONTEXT, which contains:
• RULES: short, testable guidelines derived from past evaluation mistakes.
• PROTOTYPE EXAMPLES: previously judged overlay examples with scores and brief rationales.
Use this MEMORY CONTEXT only as an internal reference to make your judgments more consistent,
geometry-aware, and physically plausible.
<Retrieved Rules> <Retrieved Prototype Images>
```

Table 1. Prompt template for memory-augmented scoring of our critique agent.

Prompt Template: Critique Agent Self-Reflection

```
You are auditing the judging rationale for 3D human mesh overlays.
You will receive:
GROUP_CONTEXT_JSON: a JSON object for a group of hypotheses that includes, for each prediction:
• the judge's score and one-sentence comment, and
• ground-truth quality scores (larger is better).
Your goal is to discover simple, testable patterns where the judge systematically over- or
under-estimates quality relative to the ground truth, and turn them into reusable RULES.

Each rule must:
• Be a single, concrete, testable sentence (e.g., \If visible thigh{torso penetration occurs,
  deduct 5-8 points.").
• Describe a condition AND its effect on the score.
• Be short, non-redundant, and general enough for future images.
• Include 2-3 short tags (e.g., ["penetration", "physical_plausibility"]).
Return STRICT JSON in the following format, and output nothing else:
{ "new_rules":  [ {
"text":  "If clear foot-ground separation is visible while the person is standing, deduct 5-8
points for poor contact realism",
"tags":  ["foot_contact", "physical_plausibility"] } ] }
```

Table 2. Prompt template for the reflective knowledge construction of our critique agent.