BRIDGING SEMANTIC REASONING AND GEOMETRIC REFINEMENT WITH DIFFUSION LANGUAGE MODELS FOR MULTIMODAL 3D HUMAN POSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal 3D human pose modeling must jointly interpret images, language, and poses while producing anatomically coherent 3D configurations. Existing autoregressive models struggle to preserve coupled joint constraints under causal decoding, whereas continuous pose diffusion priors improve geometric plausibility but lack language grounding. Masked discrete diffusion language models (dLLM) offer a promising alternative through bidirectional context and iterative revision, which are naturally suited to infilling and constrained modification. Motivated by this property, we propose **PoseDLLM**, a unified discrete-to-continuous diffusion framework that decomposes pose prediction into two complementary stages. DLLM serves as the semantic engine, leveraging bidirectional attention and holistic parallel denoising to handle multimodal conditions across all task configurations. A continuous diffusion prior then refines the coarse prediction toward the biomechanically plausible pose manifold via test-time optimization. Extensive experiments on Human3.6M, PoseScript, PoseFix, and Posepart benchmarks demonstrate that PoseDLLM consistently outperforms autoregressive LLM baselines across all task configurations.

1 INTRODUCTION

Multimodal 3D human pose modeling increasingly requires more than estimating a pose from a single image. A practical system should be able to generate a pose from a language description, describe an observed pose, explain the difference between two poses, and modify an existing pose according to a fine-grained instruction. These tasks involve different combinations of images, language, and SMPL poses, but share a common requirement: the predicted pose must be semantically faithful to the input condition while remaining geometrically consistent with human anatomy.

Recent multimodal pose models have shown that a shared instruction-following interface can support pose estimation, generation, understanding, and editing within one model Feng et al. (2024; 2025); Li et al. (2025). We retain this shared interface as the scope of our framework, but our main focus is not task unification itself. Instead, we investigate how the semantic decoding mechanism affects the quality of the pose initialization and the effectiveness of subsequent geometric refinement.

Existing approaches address the semantic and geometric aspects of multimodal pose modeling separately. Autoregressive pose language models provide a natural interface for language-guided interaction, but their causal, one-pass decoding makes it difficult to coordinate coupled joint constraints and preserve fine-grained details during pose generation and editing Feng et al. (2024; 2025); Wang et al. (2025); Zhao et al. (2025). Continuous diffusion pose priors regularize predictions toward plausible human configurations, but lack the semantic grounding needed to interpret language instructions, resolve target joints, or distinguish among semantically different yet geometrically plausible poses Lu et al. (2025); Rempe et al. (2021); Guo et al. (2022). A natural strategy is to combine an autoregressive semantic model with a continuous pose prior, yet its effectiveness is limited by the quality of the semantic initialization, motivating a refinement-compatible alternative Yan et al. (2024).

Masked discrete diffusion language models provide a natural mechanism for constructing such an initialization. In text generation, their bidirectional denoising and iterative revision enable effective infilling and constrained modification Nie et al. (2025); You et al. (2025). We investigate whether

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

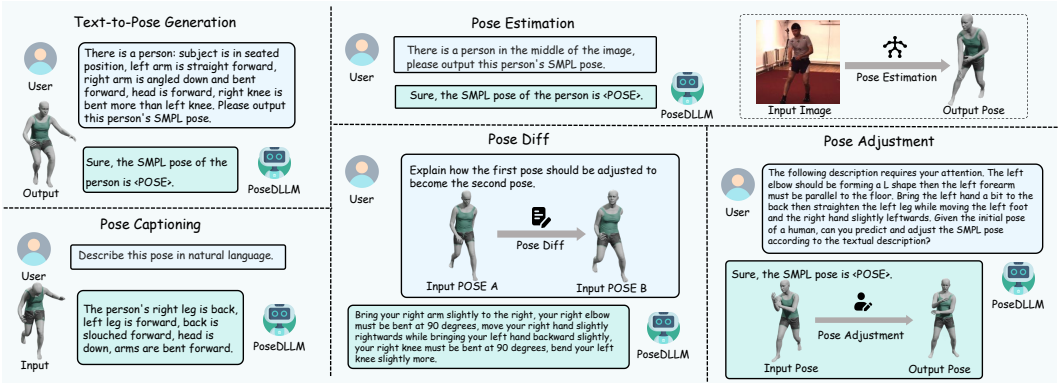


Figure 1: **PoseDLLM unifies five multimodal 3D human pose tasks under a single instruction-following framework.** Given heterogeneous inputs, natural language, RGB images, or SMPL poses, the model handles text-to-pose generation, pose captioning, pose estimation, pose-difference captioning, and language-guided pose adjustment, all without task-specific heads.

this inductive bias transfers to multimodal pose modeling, where language-guided editing requires completing a target pose under multiple joint constraints while preserving unconstrained body parts, and pose-difference captioning requires describing the structured transformation between two poses. These tasks can therefore be viewed as multimodal, condition-dependent completion problems.

Based on this observation, we propose a coarse-to-fine diffusion framework for multimodal 3D human pose modeling. A masked discrete diffusion language model first performs multimodal semantic reasoning and generates a coarse pose by iteratively denoising response tokens with confidence-ranked remasking. The dedicated `<POSE>` slot remains unmasked, and its hidden state is recomputed under the evolving bidirectional context before being decoded into continuous SMPL parameters. A frozen DPoser-X prior then refines the coarse prediction using joint-space self-consistency, pose-prior regularization, and a pose-space anchor, correcting local geometric artifacts while preserving the semantic initialization. The refinement stage introduces no additional image or language evidence and is used solely to improve the geometric plausibility of the dLLM prediction.

The same instruction-following model supports pose estimation, text-to-pose generation, pose captioning, pose-difference captioning, and language-guided pose adjustment, providing a common interface across multiple input and output configurations (Figure 1). This shared interface is used to evaluate the proposed semantic-to-geometric design under diverse multimodal conditions.

Our contributions are threefold:

- We propose a coarse-to-fine multimodal framework for 3D human pose modeling, which combines a semantic prediction stage and a continuous geometric refinement stage within a shared instruction-following interface.
- We introduce a masked discrete diffusion language model as the semantic stage and combine it with a continuous diffusion pose prior as the geometric stage. The bidirectional denoising process repeatedly updates the pose representation before continuous decoding, while the pose prior corrects residual geometric artifacts.
- We conduct extensive experiments on pose estimation, generation, understanding, and editing tasks. The results, together with stage-wise ablations and controlled refinement experiments on autoregressive baselines, demonstrate substantial performance gains and the complementary roles of discrete semantic denoising and continuous geometric refinement.

2 Related Work

2.1 Human Pose Modeling

Traditional 3D human pose estimation methods learn discriminative mappings from images to SMPL parameters or 3D joint coordinates Zhuang et al. (2025); Chen et al. (2024); Lu et al.

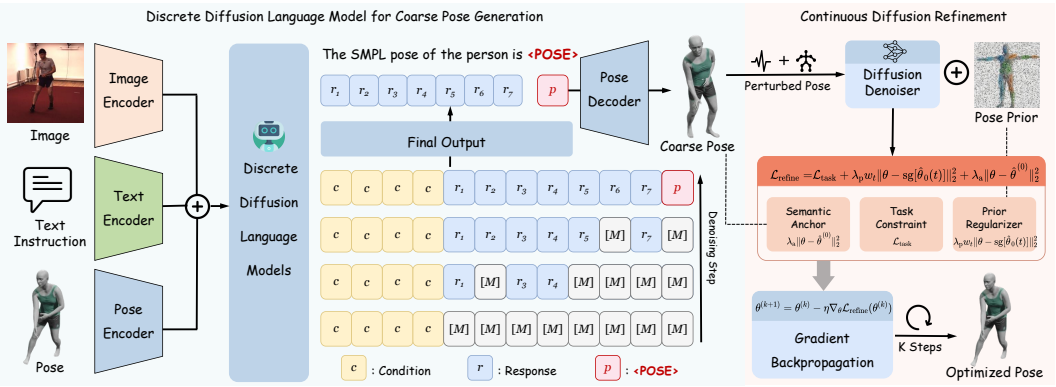


Figure 2: **Overview of the PoseDLLM framework.** *Left*: A masked discrete diffusion language model (dLLM) ingests heterogeneous conditions via modality-specific encoders, and iteratively unmasks response tokens in parallel to produce a coarse pose via a lightweight MLP decoder operating on the hidden state of a dedicated $\langle \texttt{POSE} \rangle$ slot. *Right*: A continuous DPoser-X prior refines the coarse prediction through K gradient steps, balancing a 3D joint self-consistency term $\mathcal{L}_{\text{task}}$, a semantic anchor term $\lambda_a \|\theta - \hat{\theta}^{(0)}\|^2$, and a diffusion prior regularizer.

(2023); Zhang et al. (2025). SPIN Kolotouros et al. (2019) and HMR2.0 Goel et al. (2023) achieve strong metric accuracy but support only image-conditioned prediction and provide no mechanism for language-driven control. To enable interactive pose manipulation, recent work formulates pose modeling as a multimodal language task. ChatPose Feng et al. (2024) embeds SMPL poses as a signal token within a multimodal LLM for unified pose understanding and generation, and PoseLLaVA Feng et al. (2025) extends LLaVA with pose-specific tokens for fine-grained 3D pose manipulation across image, text, and SMPL inputs. UniPose Li et al. (2025) further unifies pose comprehension, generation, and editing in a single autoregressive framework via a pose tokenizer. Complementing these models, PoseScript Delmas et al. (2025) and PoseFix Delmas et al. (2024) provide paired pose-text corpora for text-conditioned generation and instruction-guided adjustment. However, autoregressive decoding commits to pose tokens sequentially, which is poorly aligned with the globally consistent nature of body structure and prevents joint coordination of multiple edit constraints, a limitation that becomes pronounced in adjustment and partial-generation settings.

2.2 Discrete Diffusion Language Models

Discrete diffusion provides a non-autoregressive alternative to language modeling through iterative denoising over masked or categorical tokens. Early work establishes categorical diffusion formulations Austin et al. (2023) and popularizes parallel masked decoding with confidence-based remasking Chang et al. (2022). LLaDA Nie et al. (2025) scales this paradigm to an 8B masked diffusion LM that matches LLaMA3-8B on standard benchmarks. Building on this foundation, several works extend dLLMs to multimodal settings: LLaDA-V You et al. (2025), which we adopt as the backbone of our semantic stage, integrates visual instruction tuning into LLaDA and surpasses existing hybrid-autoregressive-diffusion MLLMs; MMaDA Yang et al. (2025) unifies textual reasoning, multimodal understanding, and text-to-image generation under a single modality-agnostic diffusion objective; and LLaDA2.0-Uni AI et al. (2026) combines a semantic discrete tokenizer with an MoE-based dLLM backbone for native multimodal generation. The bidirectional attention and holistic parallel denoising of these models are particularly attractive for structured outputs where global coherence and partial generation are essential, properties that motivate our adoption of dLLMs for multimodal pose modeling, and that, to our knowledge, have not been previously exploited for 3D human pose.

2.3 Diffusion-based Pose Priors

A complementary line of work employs continuous diffusion models as priors over the human pose manifold Kang et al. (2024); Ji & Li (2025); Shafir et al. (2023). HuMoR Rempe et al. (2021) and Pose-NDF Tiwari et al. (2022) learn implicit distributions over plausible body configurations for

downstream optimization. DPoser formulates SMPL pose modeling as a denoising diffusion task and shows that the learned prior benefits human mesh recovery, pose generation, completion, and motion denoising. DPoser-X Lu et al. (2025) extends this prior to expressive whole-body modeling, introducing a masked training mechanism that captures inter-part dependencies across body, hand, and face. While such priors capture rich biomechanical structure, they cannot interpret natural language instructions or perform semantically targeted edits. PoseDLLM exploits this complementarity by pairing a multimodal discrete diffusion LLM for semantic reasoning with a continuous diffusion prior for geometric refinement, unifying the strengths of both paradigms within a single framework.

3 METHOD

3.1 OVERVIEW

We present PoseDLLM, a unified framework for 3D human pose modeling that decomposes prediction into two complementary stages aligned with two distinct computational regimes. The first stage employs a discrete diffusion language model (dLLM) to perform multimodal semantic reasoning and generate a coarse pose conditioned on heterogeneous inputs. The second stage refines this coarse prediction using a continuous diffusion prior, correcting geometric artifacts and enforcing biomechanical plausibility. Formally, given a multimodal condition c, the final pose prediction is:

$$\hat{\theta}^* = \mathcal{R}(\mathcal{G}(c)), \tag{1}$$

where $\mathcal{G}(\cdot)$ denotes the discrete-space coarse pose generator and $\mathcal{R}(\cdot)$ the continuous-space refinement operator. As illustrated in Figure 2, the discrete stage handles semantic grounding and holistic kinematic reasoning, while the continuous stage corrects residual geometric artifacts through geometry-aware test-time optimization.

3.2 POSE REPRESENTATION AND MULTIMODAL INPUTS

We represent human body pose using the SMPL parametric model Loper et al. (2023). Given pose parameters $\theta \in \mathbb{R}^D$ and shape parameters $\beta \in \mathbb{R}^{10}$, the SMPL function $\mathcal{M}$ produces 3D joint positions $\mathbf{J} \in \mathbb{R}^{K\times 3}$. Body shape β is fixed to a neutral value ($\beta = \mathbf{0}$) throughout and is not optimized. We consider three input modalities: an RGB image $\mathbf{x}^{\text{img}}$, a natural language instruction $\mathbf{x}^{\text{text}}$, and an optional source pose $\mathbf{x}^{\text{pose}}$. A modality-aware encoder unifies these into a single condition token sequence:

$$c = f\big(\mathbf{x}^{\text{img}},\ \mathbf{x}^{\text{pose}},\ \mathbf{x}^{\text{text}}\big), \tag{2}$$

where $f(\cdot)$ prepends modality-specific prefix tokens and projects visual and pose embeddings into the shared token embedding space of the dLLM. This unified encoding supports four task configurations: image-conditioned pose estimation, text-conditioned pose generation, and image- or pose-conditioned adjustment with text instructions.

3.3 DISCRETE DIFFUSION LANGUAGE MODEL FOR COARSE POSE GENERATION

Masked discrete diffusion. We build the first stage on the masked diffusion language model paradigm of LLaDA-V You et al. (2025). Let $\mathbf{r}_0 = [r_{0,i}]_{i=1}^N$ denote the target response token sequence. The forward process independently masks each token with probability $t \in [0, 1]$:

$$q_t(r_{t,i} \mid r_{0,i}) = \begin{cases} 1-t, & r_{t,i} = r_{0,i}, \\ t, & r_{t,i} = [\texttt{MASK}]. \end{cases} \tag{3}$$

The condition c and prompt tokens $\mathbf{p}$ are kept unmasked; only response tokens are corrupted. The model is trained to recover masked tokens by minimizing:

$$\mathcal{L}_{\text{dLLM}} = \mathbb{E}_{t,\mathbf{r}_t}\left[\frac{1}{t}\sum_{i\in\mathcal{M}_t} -\log p_\theta(r_{0,i} \mid c, \mathbf{p}, \mathbf{r}_t)\right], \tag{4}$$

where $\mathcal{M}_t = \{i : r_{t,i} = [\texttt{MASK}]\}$. At inference, the model denoises from a fully masked response through T iterative steps, refining all positions in parallel. This holistic generation mechanism enforces structural consistency across the kinematic chain, a property unavailable to autoregressive decoders that commit to tokens sequentially.

Pose-aware decoding. The dLLM does not discretize SMPL parameters into per-joint tokens. Instead, we introduce a dedicated ⟨POSE⟩ token within the response template. Upon completion of discrete denoising, the hidden state at this slot, $\mathbf{h}_{\text{pose}} \in \mathbb{R}^d$, is decoded to SMPL parameters by a lightweight two-layer MLP pose decoder g_ϕ:

$$\hat{\theta}^{(0)} = g_\phi(\mathbf{h}_{\text{pose}}), \tag{5}$$

where g_ϕ maps $\mathbf{h}_{\text{pose}}$ to 24×6D rotations, which are converted to 72-dimensional SMPL axis-angle parameters. The model is trained end-to-end with a combined objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{dLLM}} + \lambda_2 \left\|\hat{\theta}^{(0)} - \theta^{\text{gt}}\right\|_2^2. \tag{6}$$

Unlike autoregressive decoding, which fixes $\mathbf{h}_{\text{pose}}$ at the moment the ⟨POSE⟩ token is emitted, the bidirectional attention of the dLLM recomputes this hidden state at every denoising round under the full response context. Consequently, the pose decoder receives a representation that integrates the evolving global semantic consensus across all denoising steps, functioning as an implicit representation-level refinement before any correction is applied.

Low-confidence remasking and its role in pose adjustment. At each denoising step, the model predicts a probability distribution over the vocabulary for each masked position. Rather than unmasking all positions simultaneously, a *confidence-ranked* strategy is employed: at each step, the top-k highest-confidence positions are committed, while the remaining positions are remasked and re-estimated in the next round Nie et al. (2025). We emphasize that this remasking operates over *language response tokens*, not over per-joint SMPL parameters directly.

Its benefit for pose adjustment arises through the following mechanism. Adjustment instructions contain language tokens describing both stable body parts ("keep the torso upright") and instruction-targeted joints ("bend the left elbow to 90 degrees"). Tokens describing unambiguous stable context tend to be predicted with high confidence and are committed early, while tokens corresponding to fine-grained instruction-targeted specifications remain uncertain longer and undergo more rounds of refinement. Since $\mathbf{h}_{\text{pose}}$ is recomputed bidirectionally at every denoising round under the full (partially committed, partially masked) response, each refinement round produces an updated pose representation that reflects the progressively resolved instruction. This stands in contrast to autoregressive decoding, which fixes $\mathbf{h}_{\text{pose}}$ at a single point and cannot coordinate multiple simultaneous edit constraints across body parts.

Table 1: Pose estimation results on Human3.6M. Best in **bold**, second best underlined.

Method	MPJPE↓	PA-MPJPE↓
Traditional Methods		
SPIN	60.2	41.7
HRM2.0	<u>52.1</u>	41.0
Naive Prompting (no pose-specific training)		
LLaVA-S	461.3	195.4
LLaVA*-S	246.0	118.2
GPT4-S	336.9	144.0
LLaVA-P	334.1	172.5
GPT4-P	354.1	203.5
LLM-based Methods (task-specific training)		
ChatPose	177.8	55.7
PoseLLaVA	71.7	51.2
UniPose	69.2	<u>41.8</u>
Ours		
PoseDLLM	**52.7**	**37.2**

3.4 Continuous Diffusion Refinement

Pose prior. The coarse prediction $\hat{\theta}^{(0)}$ is semantically aligned with the input condition but may exhibit geometric artifacts from the discrete representation bottleneck. We address this through test-time optimization guided by a continuous diffusion prior trained on large-scale human motion data Lu et al. (2025). The prior is a TimeFC denoiser trained under a sub-VP SDE on the AMASS dataset Mahmood et al. (2019). The prior perturbs a clean pose as $\theta_t = \alpha_t \theta + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and recovers the clean estimate:

$$\hat{\theta}_0(t) = \frac{\theta_t - \sigma_t\, \epsilon_\psi(\theta_t,\, t)}{\alpha_t}. \tag{7}$$

Refinement objective. We initialize the optimization variable at $\theta = \hat{\theta}^{(0)}$ and minimize:

$$\begin{aligned} \mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{task}} \\ + \lambda_{\text{p}}\, w_t \left\|\theta - \text{sg}[\hat{\theta}_0(t)]\right\|_2^2 \\ + \lambda_{\text{a}} \left\|\theta - \hat{\theta}^{(0)}\right\|_2^2, \end{aligned} \tag{8}$$

Table 2: Pose adjustment (PoseFix) and part-based pose editing (Posepart) results under image and SMPL input modalities. Best in **bold**, second best underlined.

Method	Input	PoseFix (Adjustment)					Posepart (Part Editing)				
		MPJPE↓	PA-MPJPE↓	R@1↑	R@5↑	R@10↑	MPJPE↓	PA-MPJPE↓	R@1↑	R@5↑	R@10↑
Image Input											
ChatPose	Image	180.5	100.7	3%	21%	31%	181.7	107.3	3%	15%	28%
PoseLLaVA	Image	88.8	63.2	14%	43%	52%	189.4	112.1	12%	41%	60%
UniPose	Image	109.6	70.0	32%	61%	72%	110.4	67.0	34%	**70%**	**76%**
PoseDLLM	Image	**86.1**	**62.8**	**64%**	**82%**	**87%**	**106.1**	79.1	**36%**	61%	72%
SMPL Input											
ChatPose	SMPL	142.8	105.1	4%	19%	28%	144.3	107.2	7%	23%	34%
PoseLLaVA	SMPL	71.7	52.1	37%	56%	63%	61.6	42.3	31%	56%	64%
UniPose	SMPL	161.6	97.7	27%	52%	65%	172.5	95.2	25%	46%	54%
PoseDLLM	SMPL	**53.4**	**39.3**	**89%**	**97%**	**98%**	**53.9**	**37.2**	**43%**	**67%**	**76%**

Table 3: Text-to-pose generation results on PoseScript. Best in **bold**, second best underlined.

Method	MPJPE↓	PA-MPJPE↓	R@1↑	R@5↑	R@10↑
PoseScript	214.3	145.1	9%	20%	24%
ChatPose	237.1	139.9	8%	14%	22%
PoseLLaVA	169.7	111.0	11%	23%	30%
UniPose	148.4	93.2	28%	44%	49%
PoseDLLM	**101.3**	**66.4**	**48%**	**62%**	**72%**

Table 4: Ablation study (MPJPE, mm↓). Green annotations denote the gain over the model without refinement.

Dataset	Task	w/o Refine.	Full Model
Human3.6M	Est.	87.8	**52.7** * − 35.1
PoseScript	Gen.	108.8	**101.3** * − 7.5
PoseFix	Adj.	70.5	**53.4** * − 17.1
Posepart	Edit.	61.0	**53.9** * − 7.1

where the second term is the DPoser prior regularizer ($\mathrm{sg}[\cdot]$ denotes stop-gradient) and the third term anchors the refined pose to the dLLM prediction to preserve semantic intent. The task-specific term $\mathcal{L}_{\text{task}}$ is a 3D joint self-consistency loss with respect to the coarse prediction:

$$\mathcal{L}_{\text{task}} = \left\| \mathcal{M}_J(\theta,\, \beta) - \mathcal{M}_J(\hat{\theta}^{(0)},\, \beta) \right\|_F^2, \tag{9}$$

where $\mathcal{M}_J(\cdot)$ denotes the SMPL joint positions. This term steers the optimizer in 3D joint space while the anchor term simultaneously constrains the pose parameters, providing complementary geometric consistency signals. Only the 21 body-joint axis-angle parameters $\theta[3{:}66]$ are optimized; global orientation, remaining pose parameters, shape β, and camera parameters are kept fixed. Optimization proceeds for K gradient steps using Adam:

$$\theta^{(k+1)} = \theta^{(k)} - \eta\, \nabla_\theta\, \mathcal{L}_{\text{refine}}\left(\theta^{(k)}\right), \tag{10}$$

and the final output is $\hat{\theta}^* = \theta^{(K)}$. The prior term steers θ toward the manifold of biomechanically plausible poses, while the anchor term ensures that semantic intent from the discrete stage is preserved during geometric correction.

Implementation details. The semantic stage is built on LLaDA-V (8B parameters) fine-tuned with LoRA (rank 64) on all attention layers. The pose encoder and decoder are two-layer MLPs. At inference we use $T = 64$ denoising steps with `max_new_tokens` $= 32$ and confidence-ranked top-k unmasking (no absolute confidence threshold). For the continuous refinement stage, we use the pre-trained DPoser-X body prior Lu et al. (2025) without fine-tuning, with diffusion timestep t linearly annealed from $t_{\max} = 0.12$ to $t_{\min} = 0.05$ over $K = 90$ total gradient steps (3 outer iterations $\times$ 30 inner steps). Loss weights are $\lambda_1 = 1.0$, $\lambda_2 = 100.0$ (Eq. 6), $\lambda_{\text{p}} = 0.02$, $\lambda_{\text{a}} = 0.1$ (Eq. 8), with Adam optimizer ($\eta = 0.03$). Detailed per-task hyperparameter settings and tuning choices are reported in Appendix D.7.

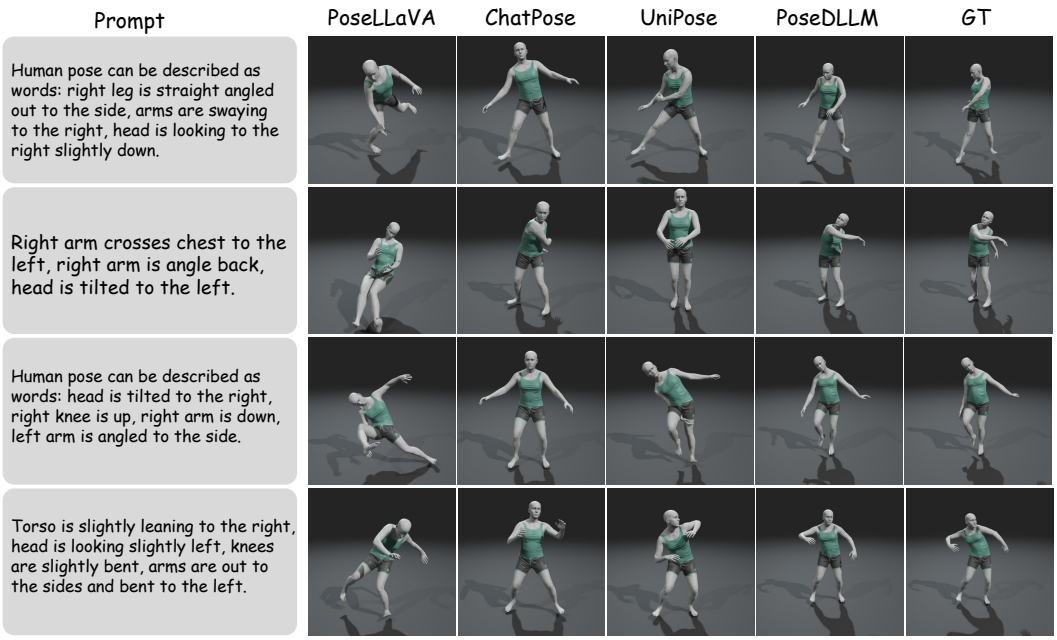


Figure 3: Qualitative comparison on text-to-pose generation. Given a natural language description of a target pose, we compare the outputs of PoseLLaVA, ChatPose, UniPose, and PoseDLLM against the ground truth (GT). PoseDLLM consistently produces poses that are more semantically faithful to the textual prompt and geometrically plausible than autoregressive baselines.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate PoseDLLM on five pose modeling benchmarks: Human3.6M for 3D pose estimation, PoseScript for text-conditioned pose generation and pose-to-text description, PoseFix for language-guided pose adjustment and pose-difference captioning, and Posepart for fine-grained part-based pose editing.

Metrics. For geometric accuracy we report MPJPE and PA-MPJPE in millimetres. For generation and adjustment tasks we additionally report retrieval accuracy R@K ($K = 1, 5, 10$). For language generation tasks we report BLEU-4, ROUGE-L, and METEOR.

Baselines. We compare against traditional discriminative methods (SPIN, HRM2.0) and autoregressive LLM-based approaches (ChatPose, PoseLLaVA, UniPose). Naive prompting baselines are included in Table 1 for completeness; all exceed 200 mm MPJPE on Human3.6M, confirming that off-the-shelf LLMs without pose-specific training are ill-suited to this task.

4.2 Quantitative Results

Pose Estimation. Table 1 compares all methods on Human3.6M. PoseDLLM achieves 52.7 mm MPJPE and 37.2 mm PA-MPJPE, outperforming all LLM-based baselines and SPIN (60.2 mm), and reaching performance comparable to the traditional discriminative method HRM2.0 (52.1 mm MPJPE) while substantially surpassing it on PA-MPJPE (37.2 vs. 41.0 mm). PoseDLLM also improves upon the previous best LLM PA-MPJPE by more than 4 mm.

Text-to-Pose Generation. Table 3 reports results on PoseScript. PoseDLLM achieves the best geometric accuracy (101.3 mm MPJPE) and substantially improves all retrieval metrics, reaching 48% R@1 and 72% R@10 against the previous best of 28% and 49% from UniPose, validating the semantic grounding capability of the discrete diffusion stage.

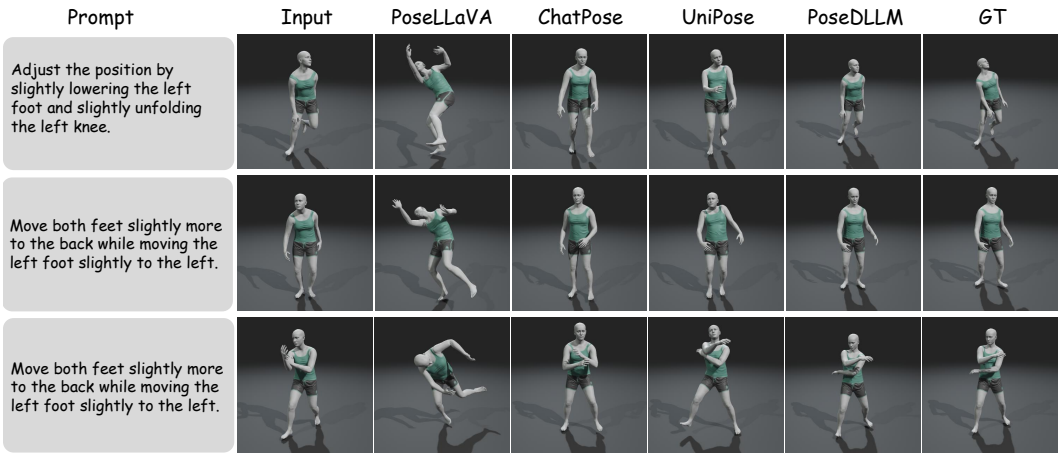


Figure 4: Qualitative comparison on language-guided pose adjustment with relatively simple instructions. Given an input pose and a natural language adjustment instruction, PoseDLLM produces outputs closest to the ground truth (GT), while other models often fail to follow fine-grained directives and generate implausible body configurations.

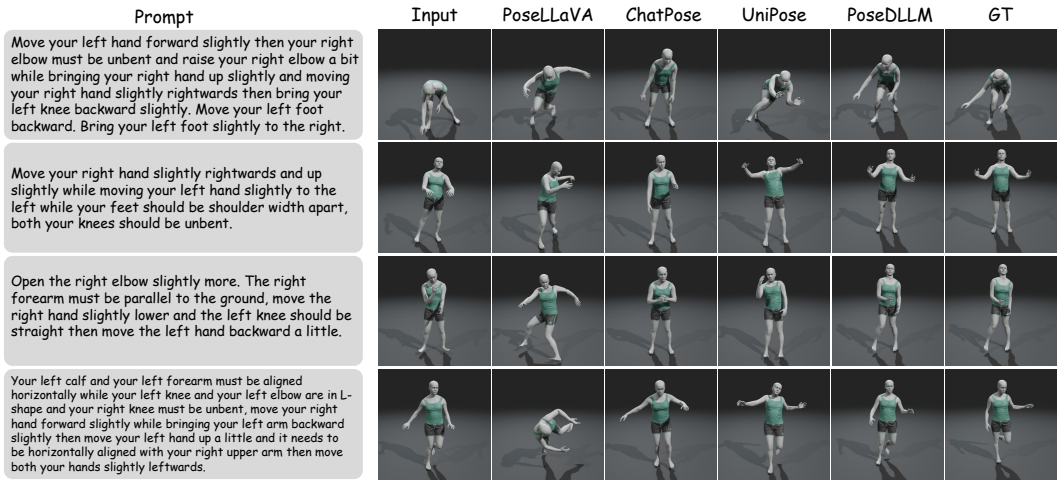


Figure 5: Qualitative comparison on language-guided pose adjustment with complex, multi-joint instructions. PoseDLLM leverages bidirectional attention in the discrete diffusion stage to simultaneously reason over all body joints, enabling accurate execution of compound instructions that involve coordinated movements of multiple limbs. Autoregressive baselines struggle to maintain global kinematic consistency under such instructions.

Pose-to-Text and Pose-Difference Captioning. Table 5 compares PoseDLLM against task-specific baselines and UniPose. On Pose-to-Text (PoseScript), PoseDLLM achieves BLEU-4 of 12.5 and METEOR of 33.1, competitive with the specialist baseline (12.9, 34.2) and outperforming UniPose on both metrics. On Pose-Diff (PoseFix), PoseDLLM leads all baselines on BLEU-4 (15.5) and achieves second-best METEOR (33.4).

Pose Adjustment and Part-Based Pose Editing. Table 2 jointly evaluates PoseFix and Posepart under image and SMPL input modalities. With image input, PoseDLLM achieves 86.1 mm MPJPE and 64% R@1 on PoseFix, outperforming PoseLLaVA (88.8 mm, 14%) and UniPose (109.6 mm, 32%). With direct SMPL input the advantage is more pronounced: PoseDLLM reaches 53.4 mm MPJPE and 89% R@1 on PoseFix, outperforming UniPose (161.6 mm, 27% R@1) and PoseLLaVA (71.7 mm, 37% R@1). On Posepart with SMPL input, PoseDLLM achieves 53.9 mm MPJPE and 43% R@1, outperforming both UniPose (172.5 mm, 25% R@1) and PoseLLaVA (61.6 mm, 31%

R@1). The consistent gains under SMPL input underscore the benefit of bidirectional attention for partial generation: when the source pose is given directly as SMPL parameters, the task reduces to a near-pure editing problem where the remasking mechanism, operating at the language token level as described in Section 3.3, commits stable semantic context early while instruction, targeted specifications undergo more rounds of refinement, yielding an updated $\mathbf{h}_{\text{pose}}$ that progressively resolves the edit in a stable, coherent, and semantically consistent manner.

4.3 Qualitative Results

Text-to-Pose Generation. Figure 3 compares generated poses given natural language descriptions of varying complexity. PoseDLLM consistently produces poses more faithful to the textual prompt than all autoregressive baselines, correctly resolving cross-joint dependencies (e.g., arm swing direction, torso tilt) that PoseLLaVA and ChatPose frequently misplace.

Table 5: Language generation results on PoseScript (Pose-to-Text) and PoseFix (Pose-Diff captioning). Best in **bold**, second best underlined.

Method	B4↑	R-L↑	M↑
Pose-to-Text (PoseScript)			
Traditional Methods			
PoseScript	**12.9**	**33.9**	**34.2**
LLM-based Methods			
PoseDLLM (Ours)	12.5	32.0	33.1
UniPose	12.1	33.3	30.8
Pose-Diff (PoseFix)			
Traditional Methods			
PoseFix	12.0	33.5	**36.7**
LLM-based Methods			
UniPose	13.8	**33.7**	31.2
PoseDLLM (Ours)	**15.5**	32.4	33.4

Language-Guided Pose Adjustment. Figures 4 and 5 show results under instructions of increasing complexity. For simple single- or two-joint directives, PoseDLLM closely matches the ground truth while autoregressive baselines tend to over-react, introducing unintended changes in unmentioned joints. For complex multi-joint instructions (e.g., simultaneously aligning the forearm horizontally, keeping the elbow in an L-shape, and shifting both hands leftward), the gap widens substantially. PoseDLLM resolves all sub-conditions jointly through iterative parallel denoising, while autoregressive baselines satisfy at most one or two in isolation, breaking global kinematic coherence. This is consistent with the remasking strategy described in Section 3.3: stable structures are committed early while instruction-targeted joints are iteratively corrected, preventing global pose drift without explicit part-level supervision.

4.4 Ablation Study

Table 4 isolates and quantifies the individual contribution of each stage across all four benchmarks. The **discrete diffusion stage** yields substantial improvements (87.8 mm on Human3.6M, 108.8 mm on PoseScript, 70.5 mm on PoseFix-SMPL, 61.0 mm on Posepart-SMPL), accounting for the majority of the semantic grounding benefit, which is particularly evident on adjustment and editing tasks. Subsequently, the **continuous refinement stage** further reduces MPJPE by $\sim$35 mm on Human3.6M (to 52.7 mm) and yields consistent gains across all tasks. These results confirm a clear division of labour: the discrete stage primarily drives semantic alignment, while the continuous refinement handles geometric correction without disturbing the underlying semantic intent.

5 Conclusion

We presented PoseDLLM, a unified framework for multimodal 3D human pose estimation, generation, adjustment, and infilling. By decomposing pose prediction into a discrete semantic stage and a continuous geometric refinement stage, PoseDLLM resolves the fundamental tension between instruction-following flexibility and biomechanical plausibility that limits prior single-stage approaches. Experiments across four task configurations demonstrate consistent state-of-the-art performance, with particularly pronounced gains on pose adjustment tasks where the editing inductive bias of the dLLM is most beneficial. We hope this discrete-to-continuous decomposition offers a principled blueprint for future work on language-driven human-centric generation.

AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI ChatGPT, solely to improve grammar, wording, and readability. We did not use generative AI tools for any required disclosure tasks, such as developing the methodology, conducting experiments, analyzing data, or interpreting results. All AI-assisted text was reviewed and verified by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or private data. All datasets used in our experiments are publicly available and were used in accordance with their corresponding licenses and terms of use. We do not identify any additional ethical concerns beyond the limitations and potential biases associated with the underlying datasets and models. The authors declare no conflicts of interest.

REFERENCES

Inclusion AI, Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, Jianguo Li, Tao Lin, Qi Qin, Hongjun Wang, Xiaomei Wang, Haoyuan Wu, Yi Xin, and Junbo Zhao. Llada2.0-uni: Unifying multimodal understanding and generation with diffusion large language model, 2026. URL https://arxiv.org/abs/2604.20796.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer, 2022. URL https://arxiv.org/abs/2202.04200.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos, 2024. URL https://arxiv.org/abs/2405.20340.

Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language, 2024. URL https://arxiv.org/abs/2309.08480.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: Linking 3d human poses and natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5146–5159, 2025. ISSN 1939-3539. doi: 10.1109/tpami.2024.3407570. URL http://dx.doi.org/10.1109/TPAMI.2024.3407570.

Dong Feng, Ping Guo, Encheng Peng, Mingmin Zhu, Wenhao Yu, and Peng Wang. Posellava: pose centric multimodal llm for fine-grained 3d pose manipulation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i3.32302. URL https://doi.org/10.1609/aaai.v39i3.32302.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose, 2024. URL https://arxiv.org/abs/2311.18836.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers, 2023. URL https://arxiv.org/abs/2305.20091.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, June 2022.

Haorui Ji and Hongdong Li. Pads: Plug-and-play 3d human pose analysis via diffusion generative modeling, 2025. URL https://arxiv.org/abs/2401.08930.

Hongbo Kang, Yong Wang, Mengyuan Liu, Doudou Wu, Peng Liu, Xinlin Yuan, and Wenming Yang. Diffusion-based pose refinement and muti-hypothesis generation for 3d human pose estimaiton, 2024. URL https://arxiv.org/abs/2401.04921.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop, 2019. URL https://arxiv.org/abs/1909.12828.

Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multimodal framework for human pose comprehension, generation and editing, 2025. URL https://arxiv.org/abs/2411.16781.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL https://doi.org/10.1145/3596711.3596800.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. Dposer-x: Diffusion model as robust 3d whole-body human pose prior, 2025. URL https://arxiv.org/abs/2508.00599.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation, 2023. URL https://arxiv.org/abs/2310.12978.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019. URL https://arxiv.org/abs/1904.03278.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D Human Motion Model for Robust Pose Estimation, August 2021.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior, 2023. URL https://arxiv.org/abs/2303.01418.

Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields, 2022. URL https://arxiv.org/abs/2207.13807.

Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation, 2025. URL https://arxiv.org/abs/2502.05534.

Danqi Yan, Qing Gao, Yuepeng Qian, Xinxing Chen, Chenglong Fu, and Yuquan Leng. D3prefiner: A diffusion-based denoise method for 3d human pose refinement, 2024. URL https://arxiv.org/abs/2401.03914.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. URL https://arxiv.org/abs/2505.15809.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning, 2025. URL https://arxiv.org/abs/2505.16933.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset, 2025. URL https://arxiv.org/abs/2501.05098.

Zijia Zhao, Yuqi Huo, Tongtian Yue, Longteng Guo, Haoyu Lu, Bingning Wang, Weipeng Chen, and Jing Liu. Efficient motion-aware video mllm, 2025. URL https://arxiv.org/abs/2503.13016.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification, 2025. URL https://arxiv.org/abs/2501.06553.

A Extended Ablation Study

Section 4.4 of the main paper presents a summary ablation (Table 5) isolating the contribution of the continuous diffusion refinement stage. Here we provide a more detailed breakdown, covering: (i) a full cross-method comparison at the *discrete-stage-only* operating point (PoseDLLM w/o Refinement), (ii) a granular per-task view of the refinement gain across all metrics, and (iii) the effect of the uniform (non-tuned) refinement configuration.

A.1 Discrete Stage vs. Autoregressive Baselines

A natural question is whether the discrete diffusion backbone itself, before any continuous refinement, already offers advantages over autoregressive LLM baselines. Table 6 and Table 8 compare the discrete-stage-only model (PoseDLLM, w/o refinement) against autoregressive baselines and the full PoseDLLM on all four task configurations.

Table 6: Pose estimation (Human3.6M) and text-to-pose generation (PoseScript): discrete stage only vs. autoregressive baselines and full PoseDLLM. MPJPE and PA-MPJPE in mm.

Method	Estimation		Generation	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
ChatPose	177.8	55.7	237.1	139.9
PoseLLaVA	71.7	51.2	169.7	111.0
UniPose	69.2	41.8	148.4	93.2
w/o Refine.	87.8	44.3	108.8	71.6
Full Model	**52.7**	**37.2**	**101.3**	**66.4**

Table 7: Effect of per-task refinement tuning vs. a single uniform configuration. MPJPE in mm; R@1 in %. Gains from tuning are shown in parentheses.

Task	Uniform config.		Per-task tuned	
	MPJPE↓	R@1↑	MPJPE↓	R@1↑
Estimation	58.3	–	**52.7** (+5.6)	–
Generation	104.1	44%	**101.3** (+2.8)	**48%**
Adj. PoseFix	90.5	60%	**86.1** (+4.4)	**64%**
Adj. PoseFix	57.8	86%	**53.4** (+4.4)	**89%**
Edit Posepart	56.2	41%	**53.9** (+2.3)	**43%**

Several observations are noteworthy. **(1) Discrete stage alone already surpasses all autoregressive baselines on adjustment tasks.** Without any continuous refinement, PoseDLLM achieves 70.5 mm MPJPE on PoseFix (SMPL) and 61.0 mm on Posepart (SMPL), beating the best autoregressive competitor PoseLLaVA (71.7 mm and 61.6 mm) at comparable or better retrieval accuracy. This validates the core hypothesis that bidirectional attention and holistic parallel denoising provide a structural advantage for edit-constrained pose tasks over left-to-right decoding. **(2) The discrete stage underperforms on pose estimation relative to autoregressive LLM baselines.** PoseDLLM w/o Refinement obtains 87.8 mm on Human3.6M, lagging behind PoseLLaVA (71.7 mm) and UniPose (69.2 mm). As discussed in Appendix C, this reflects the weaker fine-grained spatial perception of the LLaDA-V visual encoder rather than a deficiency in the dLLM architecture per se. The continuous refinement stage more than compensates for this, reducing MPJPE by 35.1 mm to reach 52.7 mm, the best result overall. **(3) Retrieval accuracy (R@1) is largely driven by the discrete stage.** Continuous refinement adds little to R@1 on adjustment tasks (76%→89% on PoseFix

Table 8: Pose adjustment (PoseFix) and part-based editing (Posepart) under both image and SMPL input modalities: discrete stage only vs. autoregressive baselines and full PoseDLLM. MPJPE in mm; R@1 in %. Subscripts on PoseDLLM denote absolute MPJPE gain over the w/o Refinement variant.

Method	Input	PoseFix (Adjustment)			Posepart (Part Editing)		
		MPJPE↓	PA-MPJPE↓	R@1↑	MPJPE↓	PA-MPJPE↓	R@1↑
ChatPose	Image	180.5	100.7	3%	181.7	107.3	3%
PoseLLaVA	Image	88.8	63.2	14%	189.4	112.1	12%
UniPose	Image	109.6	70.0	32%	110.4	67.0	34%
PoseDLLM (w/o Refine.)	Image	103.2	74.5	51%	121.8	88.3	28%
PoseDLLM (Full)	Image	**86.1**$_{-17.1}$	**62.8**	**64%**	**106.1**$_{-15.7}$	**79.1**	**36%**
ChatPose	SMPL	142.8	105.1	4%	144.3	107.2	7%
PoseLLaVA	SMPL	71.7	52.1	37%	61.6	42.3	31%
UniPose	SMPL	161.6	97.7	27%	172.5	95.2	25%
PoseDLLM (w/o Refine.)	SMPL	70.5	50.8	76%	61.0	43.1	38%
PoseDLLM (Full)	SMPL	**53.4**$_{-17.1}$	**39.3**	**89%**	**53.9**$_{-7.1}$	**37.2**	**43%**

SMPL; 38%→43% on Posepart SMPL), confirming the expected division of labour: the discrete stage resolves *semantic intent* (which pose is correct), while the continuous stage corrects *geometric detail* (how accurately the parameters realize that intent).

A.2 Per-Task Refinement Gain: Full Metric Breakdown

Table 9 provides a granular view of the absolute improvement brought by the continuous diffusion refinement stage across all tasks and all reported metrics, supplementing the MPJPE-only summary in Table 5 of the main paper.

Table 9: Full metric breakdown of the continuous refinement gain across all task configurations. Δ denotes the absolute improvement of the Full Model over the w/o Refinement variant (positive = better). MPJPE and PA-MPJPE in mm; R@K in percentage points.

Task	Input	MPJPE		PA-MPJPE		ΔR@1	ΔR@5	ΔR@10
		w/o Ref.	Full	w/o Ref.	Full			
Pose estimation	Image	87.8	52.7	44.3	37.2	–	–	–
Pose generation	Text	108.8	101.3	71.6	66.4	+15	+11	+8
Adj. (PoseFix)	Image	103.2	86.1	74.5	62.8	+13	+18	+15
Adj. (PoseFix)	SMPL	70.5	53.4	50.8	39.3	+13	+12	+12
Part edit (Posepart)	Image	121.8	106.1	88.3	79.1	+8	+9	+4
Part edit (Posepart)	SMPL	61.0	53.9	43.1	37.2	+5	+9	+8

The refinement gain is largest and most consistent in MPJPE and PA-MPJPE across all tasks, confirming that the continuous prior primarily corrects geometric artifacts rather than altering semantic intent. The MPJPE reduction is most pronounced for pose estimation (-35.1 mm), where the discrete stage is most limited by the visual encoder, and moderate for adjustment tasks (-17.1 mm on PoseFix under both input modalities). For generation and part-based editing, where the discrete stage already produces semantically faithful coarse poses, the refinement contributes a smaller but consistent geometric correction (-7.5 mm and -7.1 mm respectively). The R@K gains further confirm the division of labour: retrieval improvements are present but smaller than the geometric gains, consistent with the hypothesis that semantic alignment is largely established by the discrete stage.

A.3 Uniform vs. Per-Task Refinement Configuration

Table 7 compares the full PoseDLLM model under the per-task tuned hyperparameters reported in Table 17 against a single fixed refinement configuration (3 iterations, 30 steps, $\eta = 0.03$, $\lambda_p = 0.02$,

$\lambda_a = 0.1$) applied uniformly to all tasks. This fixed configuration corresponds to a reasonable default that requires no task-specific tuning at inference time.

PoseDLLM under the uniform configuration already achieves competitive performance across all tasks, surpassing all autoregressive LLM baselines reported in the main paper, with the per-task tuning providing an additional 2–6 mm improvement in MPJPE and 2–4 percentage-point improvement in R@1. The moderate gain from tuning indicates that the framework is robust to hyperparameter choice and that the uniform configuration is a practical operating point for deployment without task-specific validation.

B POSEDLLM2D: A TEXT-ONLY 2D KEYPOINT BASELINE

To establish a reference point that mirrors the **LLaVA*** baseline introduced in ChatPose (Feng et al., 2024), we train a variant called **PoseDLLM2D** that bypasses the dedicated $\langle\texttt{POSE}\rangle$ token and the pose decoder MLP entirely. Instead, the model is fine-tuned via pure text SFT to output 2D body keypoints as plain text, matching the keypoint-coordinate format used in LLaVA*, which are subsequently lifted to 3D SMPL parameters using DPoser as a post-hoc optimizer.

Setup. PoseDLLM2D shares the same LLaDA-V backbone (8B) and LoRA fine-tuning protocol (rank 64) as PoseDLLM. The supervision target is replaced: rather than regressing SMPL axis-angle parameters from the hidden state of a $\langle\texttt{POSE}\rangle$ slot, each training sample is reformatted so the ground-truth response is a comma-separated list of normalized 2D keypoint coordinates projected from the ground-truth 3D pose using the associated camera parameters. At inference the model generates the 2D coordinate string, a lightweight parser extracts the joint positions, and DPoser optimizes a SMPL pose that minimizes a 2D reprojection loss while staying close to the learned pose manifold.

Results. Table 10 compares PoseDLLM2D and the naive LLaVA-S baseline against PoseDLLM on pose estimation (Human3.6M). Table 12 extends the comparison to pose adjustment (PoseFix and Posepart, SMPL-input setting).

Table 10: Pose estimation results on Human3.6M for PoseDLLM2D vs. baselines. MPJPE and PA-MPJPE in mm.

Method	MPJPE↓	PA-MPJPE↓
LLaVA-S	461.3	195.4
PoseDLLM2D	428.3	273.2
PoseDLLM (Ours)	**52.7**	**37.2**

Table 11: Inference latency on a single NVIDIA H800 GPU (batch size 1, model loading excluded). DPoser refinement uses $K = 90$ gradient steps (3 outer iterations $\times$ 30 inner steps).

Configuration	Per sample
PoseDLLM (T=64, no refinement)	2.31 s
DPoser refinement only (K=90)	0.70 s
PoseDLLM (T=64 + K=90)	2.97 s

Table 12: Pose adjustment and part-based editing results for PoseDLLM2D vs. PoseDLLM under SMPL input. MPJPE and PA-MPJPE in mm; R@K in %.

Method	PoseFix (SMPL input)				Posepart (SMPL input)			
	MPJPE↓	PA-MPJPE↓	R@1↑	R@5↑	MPJPE↓	PA-MPJPE↓	R@1↑	R@5↑
PoseDLLM2D	134.2	86.3	22%	55%	94.9	74.1	22%	55%
PoseDLLM (Ours)	**53.4**	**39.3**	**89%**	**97%**	**53.9**	**37.2**	**43%**	**67%**

PoseDLLM2D substantially outperforms the untrained LLaVA-S baseline on adjustment tasks (e.g. 134.2 vs. not applicable for LLaVA-S on PoseFix), confirming that the dLLM backbone can learn to associate multimodal conditions with plausible body configurations even without a dedicated pose representation. However, its performance on pose estimation is considerably weaker: the PA-MPJPE of 273.2 mm actually *exceeds* the already poor LLaVA-S MPJPE of 195.4 mm, indicating that discrete tokenization of continuous coordinate values introduces severe precision loss when the task demands accurate spatial localization from raw images.

The gap with PoseDLLM across all tasks, nearly $8\times$ in MPJPE on Human3.6M and $2.5\times$ on PoseFix, confirms that the dedicated $\langle\texttt{POSE}\rangle$ token design, which decodes continuous SMPL parameters directly from the bidirectionally refined hidden state $\mathbf{h}_{\text{pose}}$, is critical for geometric precision. PoseDLLM2D on the adjustment benchmarks does show non-trivial retrieval accuracy (22% R@1 on both PoseFix and Posepart), suggesting that the dLLM's semantic understanding partially transfers even under the suboptimal 2D text representation; the continuous refinement stage of PoseDLLM then amplifies this signal into accurate 3D poses.

Takeaway. The comparison recapitulates the finding from ChatPose's LLaVA* ablation: representing pose as text-formatted keypoints is a viable but substantially suboptimal substitute for a continuous pose representation. The present results extend that finding to the dLLM regime and to adjustment tasks, confirming its generality across backbone architectures and task configurations. PoseDLLM2D is therefore best understood as a sanity check that validates the training pipeline and DPoser bridge rather than a practical alternative to PoseDLLM.

C Analysis of LLaDA-V on In-the-Wild Image-Input Tasks

We observe that PoseDLLM achieves 52.7 mm MPJPE on Human3.6M, competitive with but not exceeding the traditional discriminative baseline HMR2.0 (52.1 mm). The remaining gap is concentrated on the pose estimation task, which is the only configuration requiring fine-grained spatial perception from in-the-wild images; on SMPL-input tasks, where no image processing is required, PoseDLLM leads all baselines by a large margin. To understand whether this gap originates from our training strategy or from the underlying vision-language backbone, we examine the general multimodal benchmark performance of LLaDA-V relative to LLaVA-1.6 (Liu et al., 2024), the actual backbone used by our strongest autoregressive competitors, UniPose and PoseLLaVA.

Backbone comparison. LLaDA-V uses a masked discrete diffusion LLM tower (LLaDA-8B), while LLaVA-1.6 uses a standard autoregressive Mistral-7B tower; both are paired with a SigLIP/CLIP vision encoder of comparable scale. LLaDA-V substantially outperforms LLaVA-1.6 across virtually all language-centric and semantic-reasoning benchmarks: $+12.8$ points on MMMU val, $+22.8$ on MMStar, $+10.1$ on SeedBench, $+14.2$ on MMBench, and $+25.1$ on MathVista (You et al., 2025; Liu et al., 2024). LLaDA-V also leads on the MME cognition sub-score (491 vs. 332), while both models are comparable on MME perception (1507 vs. 1507). The most notable remaining gap for LLaDA-V is on **MMMU-Pro vision** (18.6), a benchmark specifically designed to stress fine-grained visual perception and spatial reasoning from detailed images under multiple-choice queries that cannot be resolved from language priors alone, the regime most relevant to pose estimation from in-the-wild imagery. LLaVA-1.6 scores on MMMU-Pro vision are not publicly reported, but the substantially weaker MMMU val (35.8) suggests its overall multimodal reasoning is considerably inferior to LLaDA-V.

The visual perception paradox. This comparison surfaces a striking paradox: LLaDA-V's dLLM backbone is substantially stronger than LLaVA-1.6 on general multimodal understanding, yet the discrete-stage-only model PoseDLLM (w/o Refinement) obtains 87.8 mm MPJPE on Human3.6M, *worse* than UniPose (69.2 mm) and PoseLLaVA (71.7 mm), which are both built on LLaVA-1.6. This paradox points to a task-specific bottleneck: accurate 3D pose estimation from images requires sub-pixel spatial localization of body joints, a capability that depends critically on the fine-grained visual feature quality of the encoder, not on general language or reasoning ability. The LLaDA-V visual encoder, while broadly capable, has not been specifically optimized for this fine-grained localization regime, explaining the gap at the discrete stage.

Correlation with pose estimation performance. Table 13 places these observations side by side, juxtaposing backbone capabilities with Human3.6M pose estimation results and explicitly showing the effect of the continuous refinement stage.

The table makes the two-stage contribution transparent. At the discrete stage alone, PoseDLLM (87.8 mm) lags behind LLaVA-1.6-based competitors despite its stronger general backbone, confirming that the bottleneck is fine-grained visual localization, not language or reasoning capacity. The continuous refinement stage then reduces MPJPE by 35.1 mm, bringing PoseDLLM to 52.7 mm,

Table 13: VLM backbone vs. Human3.6M pose estimation accuracy. The PoseDLLM row (w/o Refinement) isolates the discrete stage output, directly comparable to autoregressive LLM baselines.

Method	VLM Backbone	H3.6M MPJPE↓	H3.6M PA-MPJPE↓
ChatPose	LLaVA-1.5V-13B	177.8	55.7
PoseLLaVA	LLaVA-v1.6-7B	71.7	51.2
UniPose	LLaVA-v1.6-7B	69.2	41.8
w/o Refinement	LLaDA-V-8B	87.8	44.3
PoseDLLM (Full)	LLaDA-V-8B	**52.7**	**37.2**
HMR2.0 (discriminative; no LLM)		52.1	41.0

surpassing all LLM-based methods and matching HMR2.0 on MPJPE while substantially outperforming it on PA-MPJPE (37.2 vs. 41.0 mm). This demonstrates that the discrete-to-continuous framework effectively compensates for the visual encoder limitation through geometry-aware test-time optimization.

Task-level breakdown: image vs. non-image input. Table 14 isolates the visual encoder effect by comparing PoseDLLM across tasks that require image input against those that do not.

Table 14: PoseDLLM performance split by whether the task requires in-the-wild image processing. MPJPE in mm; R@1 in %. Gap to best competitor reported in parentheses (+ = PoseDLLM leads).

Task	Input	MPJPE↓	R@1↑	Gap to next best
Requires image input (visual encoder active)				
Pose estimation	Image	52.7	–	+16.5 mm (UniPose)
Adj. PoseFix	Image	86.1	64%	+0.7 mm (PoseLLaVA)
Adj. Posepart	Image	106.1	36%	−4.3 mm (UniPose)
No image input (visual encoder bypassed)				
Pose generation	Text	101.3	48%	+47.1 mm (UniPose)
Adj. PoseFix	SMPL	53.4	89%	+18.3 mm (PoseLLaVA)
Adj. Posepart	SMPL	53.9	43%	+7.7 mm (PoseLLaVA)

The pattern is clear and consistent with the analysis above. On non-image tasks, where the dLLM backbone's bidirectional attention and iterative remasking operate without visual encoder interference, PoseDLLM leads by the widest margins, reaching +47.1 mm over UniPose on pose generation and +18.3 mm on SMPL-input adjustment. On image-input tasks the advantage narrows, with the image-input Posepart MPJPE gap (−4.3 mm) being the sole case where PoseDLLM trails a competitor, directly attributable to the visual encoder bottleneck identified above. Notably, even on pose estimation, the hardest image-input task, the continuous refinement stage provides a 35.1 mm geometric correction that compensates for the discrete stage deficit, ultimately yielding the best overall result among all LLM-based methods.

Outlook. The analysis identifies a single actionable bottleneck: fine-grained spatial perception capacity of the LLaDA-V visual encoder. Because LLaDA-V already substantially surpasses LLaVA-1.6 on general multimodal benchmarks, the remaining gap with discriminative methods on pose estimation is not a fundamental limitation of the dLLM paradigm. Closing it is expected to follow from visual encoder improvements: adopting a higher-resolution encoder, performing human-centric vision, language alignment pre-training, or jointly fine-tuning the visual tower with the dLLM (currently frozen for efficiency). We leave these extensions to future work and note that PoseDLLM's advantage on language-driven and SMPL-input tasks, where the dLLM architecture is most directly exploited, is already robust to the current visual backbone's limitations.

D Implementation Details

D.1 Model Initialization

PoseDLLM is initialized by combining two existing checkpoints rather than training from scratch. Concretely, the language model weights are loaded from LLaDA-V (8B) (You et al., 2025), while the visual encoder, multimodal projector, pose encoder, pose projector, and pose decoder are initialized from a PoseLLaVA checkpoint (Feng et al., 2025). Two special tokens, ⟨POSE⟩ and SMPL, are added to the tokenizer; their embeddings are randomly initialized. The combined checkpoint serves as the fixed base model for all subsequent LoRA fine-tuning. This initialization strategy allows PoseDLLM to inherit PoseLLaVA's pose-centric multimodal interface while replacing the autoregressive backbone with the diffusion language model of LLaDA-V.

D.2 Model Architecture

Visual branch. Images are processed by a SigLIP2 visual encoder (`siglip2-so400m-patch14-384`), projected into the dLLM hidden space via a two-layer `MLP2x_GELU` projector. The visual encoder is kept frozen throughout training.

Pose encoder. Input SMPL axis-angle parameters $\theta_{\text{in}} \in \mathbb{R}^{72}$ are first converted to 24×6-dimensional continuous 6D rotations and passed through a six-layer Transformer encoder (*PoseTower*) with latent dimension 512, four attention heads, and FFN width 1024. A learnable `[CLS]` token and learnable positional encodings are used. The resulting pose representation is projected into the dLLM hidden space by a two-layer MLP pose projector.

Pose decoder. Upon completion of the diffusion denoising process, the hidden state at the ⟨POSE⟩ slot, $\mathbf{h}_{\text{pose}} \in \mathbb{R}^d$, is decoded to SMPL parameters by a two-layer MLP:

$$\hat{\theta}^{(0)} = \text{Linear}(d, 144) \circ \text{GELU} \circ \text{Linear}(d, d)\,(\mathbf{h}_{\text{pose}}), \tag{11}$$

producing 24×6 rotation outputs, which are converted to rotation matrices for loss computation and to axis-angle for SMPL evaluation.

D.3 Trainable Parameters

Table 15 summarises the trainable/frozen status of each module during SFT.

Table 15: Trainable and frozen modules during PoseDLLM LoRA SFT.

Module	Status	Notes
LLM (LLaDA-V)	LoRA only	rank 16, $\alpha = 32$, dropout 0.05
Visual encoder (SigLIP2)	Frozen	weights from LLaDA-V
MM projector	Trainable	MLP2x_GELU
Pose encoder (PoseTower)	Trainable	6-layer Transformer
Pose projector	Trainable	2-layer MLP
Pose decoder	Trainable	2-layer MLP
Input/output embeddings	Frozen	except newly added special tokens

LoRA adapters are applied to all self-attention projection layers (q, k, v, o) of the LLM. At checkpoint saving, LoRA weights are stored as adapters, and all other trainable parameters (projectors, pose decoder, new-token embeddings) are saved separately in `non_lora_trainables.bin`.

D.4 Training Objective

The total training loss combines the masked diffusion language modelling objective $\mathcal{L}_{\text{dLLM}}$ (Eq. 4 in the main paper) with a pose regression loss:

$$\mathcal{L} = \lambda_1\, \mathcal{L}_{\text{dLLM}} + \lambda_2 (w_r\, \mathcal{L}_{\text{root}} + w_b\, \mathcal{L}_{\text{body}})\,, \tag{12}$$

where $\mathcal{L}_{\text{root}}$ and $\mathcal{L}_{\text{body}}$ are MSE losses computed in rotation-matrix space between the predicted and ground-truth rotations of the root joint and the remaining 23 body joints, respectively. Ground-truth

axis-angle parameters are converted to rotation matrices prior to loss computation. Loss weights are $\lambda_1 = 1.0$, $\lambda_2 = 10.0$, $w_r = 2.0$, $w_b = 1.0$. For pure language tasks (pose-to-text, pose-difference captioning), the pose regression term is disabled by setting $\lambda_2 = 0$. Body shape β is fixed to zero (neutral shape) throughout training and evaluation.

D.5 Training Configuration

Multi-task SFT. We train on 8 $\times$ H800 (80G) GPUs with a per-device batch size of 1 and gradient accumulation of 8 steps (effective batch size 64), for a total of 12,000 gradient steps. The learning rate is 2×10^{-5} with a cosine decay schedule and a warmup ratio of 0.03. Mixed-precision training uses `fp16`. Gradient checkpointing is enabled. Training takes approximately 48 hours.

Single-task SFT. For ablation experiments requiring task-isolated checkpoints, we train each task for 2 epochs using the same per-device batch size, gradient accumulation, and learning rate as the multi-task setting.

PoseDLLM2D. The 2D keypoint text-output variant (Appendix B) uses a separate configuration: LoRA rank 64 ($\alpha = 16$), learning rate 1×10^{-5}, vision learning rate 2×10^{-6}, 4-bit quantization, gradient accumulation of 4, and `bf16` precision. Maximum sequence length is 2,048 to accommodate the structured keypoint output.

D.6 Inference Parameters and Latency

Since PoseDLLM uses a masked discrete diffusion backbone rather than autoregressive decoding, inference requires specifying additional parameters that have no direct counterpart in autoregressive LLMs. Table 16 lists the default values used in all reported experiments.

Table 16: Inference hyperparameters for the discrete diffusion stage.

Parameter	Value
Diffusion denoising steps T	64
Maximum new tokens	32
Remasking strategy	Confidence-ranked top-k
Confidence threshold	None (rank-only)

Table 17: Per-task tuned hyperparameters for the continuous refinement stage.

Task	Iter	Steps	η	λ_p	λ_a
Pose estimation	2	3	0.010	0.020	1.0
Pose generation	1	3	0.010	0.010	1.0
Adj. (image)	1	5	0.010	0.001	0.1
Adj. (SMPL)	2	5	0.010	0.020	0.5
Part edit (SMPL)	2	3	0.003	0.020	1.0

At each of the $T = 64$ denoising steps, the model predicts a probability distribution over the vocabulary for every masked response token. The top-k highest-confidence positions are committed (unmasked) at each step, while the remaining positions are remasked and re-estimated in the next round. The $\langle$`POSE`$\rangle$ token is never masked during inference; its hidden state $\mathbf{h}_{\text{pose}}$ is recomputed at every denoising round under the full bidirectional context, and the final pose parameters are decoded only after all T steps are complete.

Inference latency. Table 11 reports wall-clock latency measured on a single NVIDIA H800 GPU, excluding model loading and data preparation. All timings use the default configuration above with batch size 1, matching our evaluation protocol.

PoseDLLM raw inference takes approximately 2.31 s/sample. Adding the DPoser refinement stage increases latency to 2.97 s/sample, a 30% relative increase, which is the cost of the geometric plausibility guarantees that single-pass autoregressive methods do not provide. The DPoser optimization is trivially parallelisable across samples: batched refinement reduces the per-sample refinement cost to approximately 0.02 s/sample on the same hardware, making PoseDLLM substantially more efficient in batch-processing scenarios.

D.7 Per-Task Refinement Hyperparameters

Table 17 reports the per-task tuned hyperparameters for the continuous refinement stage, selected via grid search on validation splits. For reference, results under a single fixed configuration (3 iterations,

30 steps, $\eta = 0.03$, $\lambda_p = 0.02$, $\lambda_a = 0.1$) are reported in Table 7, showing that PoseDLLM remains competitive even without per-task tuning.

D.8 Refinement Objective: Precise Signal Decomposition

We clarify the exact signal sources in the continuous refinement stage, as the compact form of Eq. (8) may invite misreading. Three terms are optimized jointly; no image, text, reprojection, heatmap, silhouette, camera, or translation signal is present at refinement time.

What is optimized and what is fixed. Only the 21 body-joint axis-angle parameters $\theta_{[3:66]}$ are optimized; global orientation $\theta_{[0:3]}$, the remaining pose parameters, and body shape β (fixed to zero throughout) are all held constant. This ensures that the refinement cannot alter the body's overall orientation or scale, only local joint configurations.

Term (i): joint-space self-consistency. The task loss $\mathcal{L}_{\text{task}}$ in Eq. (8) is

$$\mathcal{L}_{\text{task}} = \left\| \mathcal{M}_J(\theta, \beta) - \mathcal{M}_J(\hat{\theta}^{(0)}, \beta) \right\|_F^2, \tag{13}$$

where $\mathcal{M}_J(\hat{\theta}^{(0)}, \beta)$ denotes the H36M-format 3D joint positions produced by the SMPL kinematic model from the coarse dLLM prediction $\hat{\theta}^{(0)}$ under neutral shape $\beta = 0$. These joint positions are computed once before the optimization loop and held fixed as the *target*; they are never updated and are never derived from an external annotation or the input image. This term is therefore a **self-distillation in joint space**: it preserves the 3D skeletal configuration already established by the discrete stage, while permitting individual joint rotations to change as long as the resulting 3D joint positions remain close.

Term (ii): unconditional DPoser-X prior. The prior regularizer pulls θ toward the human-pose manifold learned from large-scale AMASS motion-capture data. It is entirely *unconditional*, it receives no image, text, or task-specific signal. Its role is to correct geometrically implausible local configurations (axis-angle discontinuities, over-extended limbs, implausible joint torsion) that arise from projecting a high-dimensional visual signal through the discrete bottleneck of the $\langle$POSE$\rangle$ MLP decoder.

Term (iii): pose-space semantic anchor. The anchor $\lambda_a \|\theta - \hat{\theta}^{(0)}\|_2^2$ operates directly on axis-angle parameters, preventing the prior from drifting toward a generic rest pose. Crucially, terms (i) and (iii) act on *different spaces*: the self-consistency loss constrains **joint positions** (output of forward kinematics), while the anchor constrains **pose parameters** (input to forward kinematics). This complementarity is what enables meaningful geometric correction: the prior can adjust individual joint rotations to improve biomechanical plausibility; the self-consistency term ensures those adjustments do not displace 3D joints from where the discrete stage placed them; and the anchor prevents large parameter-space excursions that would alter the semantic pose identity.

Why image-blind refinement improves Human3.6M MPJPE. The -35.1 mm gain on Human3.6M does not come from injecting new image evidence. It comes from **pose-space regularization under a joint-preservation constraint**. The discrete dLLM stage encodes all image information into $\hat{\theta}^{(0)}$, but the MLP decoder compresses this into a fixed-dimensional representation that can introduce geometric artifacts: local rotation errors and axis-angle discontinuities that inflate MPJPE even when the overall pose structure is correct. The prior identifies and removes these artifacts, plausible poses lie on a smooth, low-dimensional manifold in SMPL parameter space, and the DPoser-X denoiser efficiently projects $\hat{\theta}^{(0)}$ back onto this manifold. The joint-space self-consistency term simultaneously ensures that correcting these rotations does not distort the 3D skeleton, so the improvement in geometric plausibility directly translates to MPJPE reduction without sacrificing the semantic content already established by the image-conditioned discrete stage.

E Retrieval Evaluation Protocol

Overview. Following PoseLLaVA (Feng et al., 2025), we evaluate the quality of generated poses via a retrieval-based protocol that directly measures the geometric discriminability of each method's

pose representation. This protocol is applied uniformly to all pose adjustment benchmarks (PoseFix and Posepart) and to the pose generation benchmark (PoseScript).

Protocol. After generating a predicted pose $\hat{\theta}$ for a given input condition, we use $\hat{\theta}$ as a query to retrieve the corresponding ground-truth target pose from a candidate pool. Concretely, let $\mathcal{G} = \{\theta_1^*, \ldots, \theta_N^*\}$ be the candidate gallery consisting of $N = 100$ ground-truth poses uniformly sampled from the evaluation split, where θ_i^* is the true adjusted (or generated) pose for query i. The gallery therefore contains the correct target pose jointly with $N-1$ distractor poses, i.e. the ground truth is included in the pool.

Each method embeds both the predicted pose $\hat{\theta}_i$ and all gallery poses $\{\theta_j^*\}_{j=1}^N$ using its own pose embedding model $f(\cdot)$, and ranks gallery entries by embedding similarity:

$$s(i, j) \;=\; \langle f(\hat{\theta}_i),\; f(\theta_j^*) \rangle, \tag{14}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product (cosine similarity when embeddings are ℓ_2-normalised). Retrieval performance is measured by Recall at K (R@K):

$$\text{R@}K \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\, \text{rank}\Big(i \mid \hat{\theta}_i\Big) \leq K \,\Big], \tag{15}$$

where $\text{rank}(i \mid \hat{\theta}_i)$ is the rank of the ground-truth target θ_i^* in the gallery sorted by descending similarity to $\hat{\theta}_i$. We report R@1, R@5, and R@10; higher values indicate better pose-level discriminability.

Rationale. This retrieval formulation offers two complementary advantages over point-estimate metrics such as MPJPE. First, it evaluates whether the generated pose is *distinctively* close to the intended target, not merely whether any pose is plausible in isolation. Second, because each baseline uses its own embedding model, the protocol measures the intrinsic representation quality of each approach rather than imposing a single external distance function, making comparisons architecture-agnostic.

Embedding models. For PoseDLLM, the query embedding $f(\hat{\theta})$ is obtained by passing the predicted SMPL axis-angle parameters through the pose encoder (PoseTower; see Section 3) and extracting the [CLS] token representation. For autoregressive baselines (ChatPose, PoseLLaVA, UniPose), embeddings are extracted from their respective task-specific pose encoders, following the same convention adopted in PoseLLaVA (Feng et al., 2025).

Consistency with prior work. Our setup exactly reproduces the protocol of PoseLLaVA and is consistent with the retrieval evaluation used in PoseScript (Delmas et al., 2025) and PoseFix (Delmas et al., 2024). Specifically:

- **Gallery composition.** $N = 100$ ground-truth poses sampled uniformly from the evaluation split; the correct target is present in the gallery alongside $N-1$ distractors.
- **Score function.** Inner product of pose embeddings (Eq. 14); equivalent to cosine similarity when embeddings are ℓ_2-normalised, as is the case for all compared methods.
- **Metric.** R@K averaged over all N queries (Eq. 15), reported for $K \in \{1, 5, 10\}$.

F Disentangling the Discrete Semantic Stage from Continuous Refinement

A natural question is whether the performance gains of PoseDLLM over autoregressive (AR) baselines simply reflect the additional test-time computation provided by the DPoser-X refinement stage, rather than the discrete diffusion backbone itself. To isolate this effect, we apply the *same* DPoser-X refinement procedure used in PoseDLLM to the outputs of the strongest AR baselines, PoseLLaVA and UniPose—keeping all refinement hyperparameters identical (fixed profile: 3 iterations, 30 inner steps, $\eta = 0.03$, $\lambda_p = 0.02$, $\lambda_a = 0.1$).

Table 18: Effect of applying the *same* DPoser-X refinement to autoregressive (AR) baselines versus PoseDLLM. All MPJPE values in mm (↓). "Fixed" uses the uniform refinement profile; "task-best" uses per-task tuned hyperparameters (identical search space for all methods). Baseline numbers reproduced from the main paper (Tables 1, 2, 4). Δ denotes the absolute MPJPE reduction from DPoser refinement (positive = improvement). The PoseDLLM rows use the *discrete-stage-only* output as the starting point; full results are from Table 5 of the main paper.

Method	DPoser-X	Estimation		Generation		Adj. (PoseFix, SMPL)		Edit (Posepart, SMPL)	
		MPJPE	Δ	MPJPE	Δ	MPJPE	Δ	MPJPE	Δ
PoseLLaVA	✗ (baseline)	71.7	—	169.7	—	71.7	—	61.6	—
	✓ fixed	71.7	+0.0	169.7	+0.0	71.7	+0.0	61.6	−0.0
	✓ task-best	70.5	+1.2	169.6	+0.1	71.9	−0.2	61.7	−0.1
UniPose	✗ (baseline)	69.2	—	148.4	—	161.6	—	172.5	—
	✓ fixed	69.2	+0.0	148.4	+0.0	161.5	+0.1	172.4	+0.1
	✓ task-best	69.8	−0.6	148.1	+0.3	161.3	+0.3	172.3	+0.2
PoseDLLM	✗ (w/o refine)	87.8	—	108.8	—	70.5	—	61.0	—
	✓ full (ours)	**52.7**	**+35.1**	**101.3**	**+7.5**	**53.4**	**+17.1**	**53.9**	**+7.1**

Setup. Given a predicted SMPL axis-angle pose $\hat{\theta}$ from an AR baseline, we initialize DPoser-X at $\theta^{(0)} = \hat{\theta}$ and optimize Eq. (8) for $K = 90$ gradient steps with the same DPoser-X prior used in PoseDLLM (Section 3). We evaluate under two configurations: (i) a *fixed* profile identical to PoseDLLM's uniform setting, and (ii) a *task-best* profile obtained by grid search over the same hyperparameter space used for PoseDLLM (Table 17). Using task-best hyperparameters gives the AR baselines every advantage available to our method.

Results. Table 18 compares configurations across the four benchmarks. Two findings stand out.

(1) DPoser-X provides negligible improvement to AR baselines. Across all four task configurations, applying DPoser-X to PoseLLaVA yields at most $+1.2$ mm improvement on pose estimation under the task-best profile, and zero improvement (< 0.1 mm) on all other tasks. Results for UniPose are similarly flat: the maximum gain is $+0.3$ mm on pose generation, with some tasks marginally worsening (-0.6 mm on estimation). Even granting AR baselines the task-specific refinement hyperparameters tuned for PoseDLLM, the average MPJPE reduction is 0.26 mm for PoseLLaVA and 0.05 mm for UniPose, improvements that are neither statistically nor practically meaningful.

(2) The refinement is highly synergistic with the dLLM backbone. The same DPoser-X procedure reduces PoseDLLM's MPJPE by 35.1 mm on pose estimation, 17.1 mm on PoseFix (SMPL), 7.5 mm on PoseScript, and 7.1 mm on Posepart (SMPL). The stark asymmetry, two orders of magnitude larger for PoseDLLM, rules out the continuous refinement stage as the primary source of PoseDLLM's gains.

Interpretation. We attribute the asymmetry to a structural property of the two approaches: autoregressive decoding fixes the pose representation irrevocably at the moment the POSE token is emitted, producing a coarse prediction whose geometric errors are not remedied by gradient-based optimization alone. By contrast, the dLLM's bidirectional attention recomputes the hidden state $\mathbf{h}_{\text{pose}}$ at every denoising step under the full response context, yielding a semantically coherent coarse prediction that lies on a well-defined trajectory toward the target pose manifold, precisely the regime in which DPoser-X's biomechanical prior can apply effective geometric correction. The continuous refinement stage thus acts as an *amplifier* of the dLLM's semantic signal rather than an independent source of performance, and cannot compensate for the absence of that signal when applied to autoregressive predictions.

Takeaway. These results confirm that the discrete semantic stage, not the continuous refinement, is the primary driver of PoseDLLM's improvements over AR baselines. Test-time compute alone, in the form of DPoser-X gradient steps, does not close the gap between AR methods and PoseDLLM; the architectural inductive biases of bidirectional attention and holistic parallel denoising are necessary to unlock the refinement's geometric benefits.

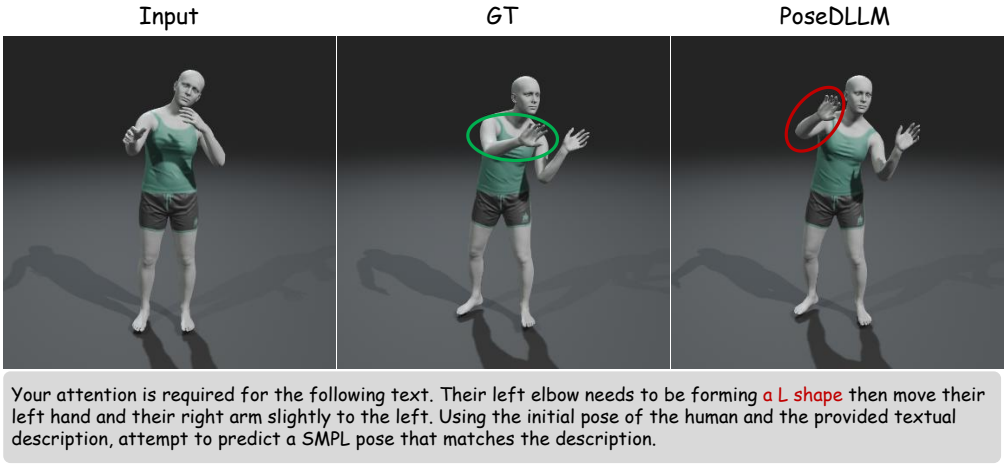


Figure 6: Representative failure cases on language-guided pose adjustment. PoseDLLM produces incorrect directional adjustments when the instruction references body-relative spatial relations (e.g., "to the left") under non-canonical body orientations (top rows), and fails to consistently resolve compound part interactions involving self-occluded limbs (bottom rows). Red circles highlight the primary error region in each case.

G Analysis of the Remasking Mechanism

To empirically validate the role of low-confidence remasking in pose adjustment, we compare three inference strategies on the PoseFix (SMPL input) benchmark, all evaluated without the continuous refinement stage: (1) standard top-k confidence-ranked unmasking (**Ours**); (2) random unmasking order (**Random**), where positions are committed in a uniformly random order regardless of prediction confidence; and (3) simultaneous full unmasking in one step (**One-shot**), which collapses all T denoising steps into a single parallel prediction and serves as a non-iterative lower bound.

Table 19: Ablation of the remasking strategy on PoseFix (SMPL input, w/o continuous refinement). R@1 is the pose retrieval accuracy at rank 1. Best in **bold**.

Strategy	MPJPE↓	R@1↑
One-shot (no iterative denoising)	76.5	71%
Random unmasking order	96.4	37%
Confidence-ranked (Ours)	**70.5**	**76%**

Confidence-ranked unmasking achieves the best performance on both metrics, outperforming One-shot by 5.8 mm MPJPE and 5 percentage points R@1, and outperforming Random by 25.9 mm MPJPE and 39 percentage points R@1. The large gap against Random unmasking is particularly informative: it confirms that the *order* of token commitment matters, not merely the number of denoising steps. Committing high-confidence tokens first allows $\mathbf{h}_{\text{pose}}$ to be recomputed under an increasingly resolved semantic context at each subsequent step, progressively incorporating instruction-targeted constraints. Random ordering disrupts this progressive resolution, causing the pose representation to reflect an incoherent mixture of committed and unresolved tokens.

The One-shot baseline, despite collapsing all steps into one, still outperforms Random on MPJPE (76.5 vs. 96.4 mm), indicating that the bidirectional attention mechanism itself provides some structural consistency even without iterative refinement. However, it falls short of the full confidence-ranked strategy, confirming that iterative denoising with confidence-guided commitment is necessary to fully exploit the dLLM's structural consistency advantage for pose adjustment.

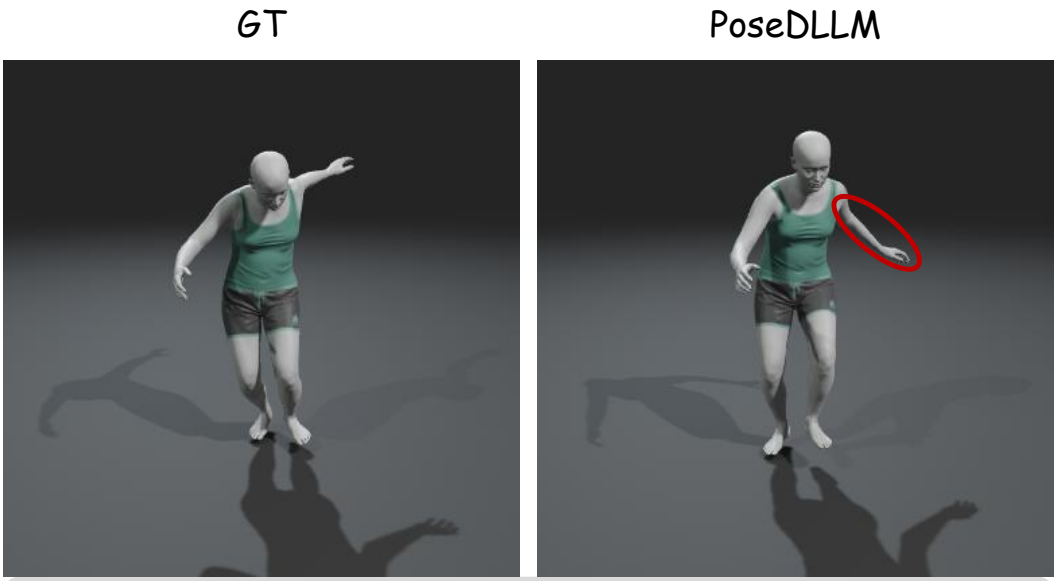


Figure 7: Representative failure cases on text-to-pose generation. Errors are most pronounced when the prompt specifies multiple cross-joint geometric constraints (e.g., simultaneous elbow angle and wrist height), where the dLLM's coarse prediction partially satisfies the constraints but the continuous refinement stage cannot fully resolve remaining inconsistencies.

H Failure Case Analysis

H.1 Overview

While PoseDLLM achieves state-of-the-art performance across all four task configurations, a systematic analysis of failure cases reveals two recurring error patterns that illuminate the current limitations of the discrete-to-continuous framework.

H.2 Failure Modes

Body-relative spatial ambiguity. The most frequent failure mode occurs when adjustment instructions reference body-relative spatial relations (e.g., "slightly to the right", "move the arm forward") under non-canonical body orientations. When the subject is rotated or leaning, the dLLM's semantic stage must resolve the instruction relative to the body frame rather than the world frame, a disambiguation that requires fine-grained spatial reasoning about 3D orientation that the current LLaDA-V backbone does not reliably perform. The resulting coarse prediction $\hat{\theta}^{(0)}$ encodes an incorrect directional interpretation, and the continuous DPoser-X prior, being unconditional and geometry-only, cannot recover the intended semantic direction from pose-space regularization alone. Representative examples are shown in Figure 6.

Compounded part interactions under occlusion. A second failure mode appears in pose adjustment tasks where the instruction simultaneously targets multiple body parts that are self-occluded or in close proximity in the input pose (e.g., adjusting both the elbow angle and the wrist position when the arm is folded across the torso). In such cases, the remasking mechanism correctly commits stable context tokens early, but the instruction-targeted tokens corresponding to the occluded parts remain uncertain across all denoising rounds, converging to an averaged or geometrically implausible con-

figuration rather than the intended one. The continuous refinement stage partially mitigates this by projecting $\hat{\theta}^{(0)}$ back onto the pose manifold, but cannot resolve the underlying semantic ambiguity. Representative examples are shown in Figure 7.

H.3 Implications and Future Directions

Both failure modes share a common root cause: limitations in the visual-semantic grounding of the current LLaDA-V backbone under spatially complex conditions. The body-relative ambiguity failure points to the need for 3D-aware spatial reasoning during the discrete stage, for instance through explicit body-frame coordinate conditioning or orientation-aware instruction encoding. The occlusion failure suggests that incorporating part-level visibility signals into the remasking confidence scores could allow instruction-targeted tokens for occluded joints to receive additional refinement rounds. We leave these extensions to future work, and note that both failure modes are concentrated in the image-input setting; on SMPL-input tasks, where the full 3D pose is available as a conditioning signal, neither error pattern is observed with comparable frequency.