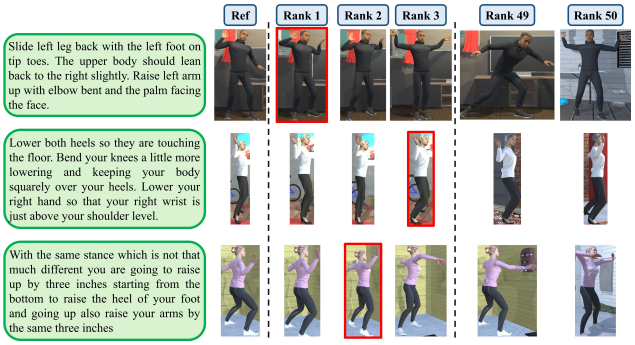

Figure 1. **Additional CPR results on FIXMYPOSE.** Ranked retrieval results are shown for a given reference image (Ref) and a transitional pose description (green box) as the query. Red boxes indicate the correct target images.

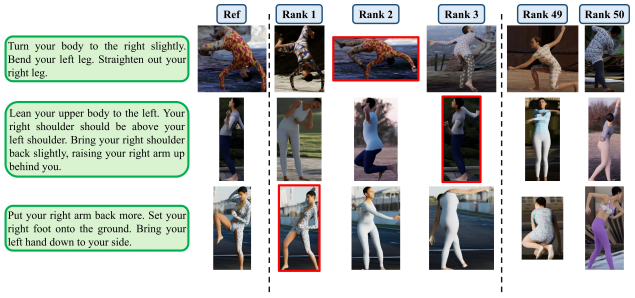

Figure 2. **Additional CPR results on PoseFixCPR.** Ranked retrieval results are shown for a given reference image (Ref) and a transitional pose description (green box) as the query. Red boxes indicate the correct target images.

1. Additional Results

Qualitative Results. Similar to Figure 4 in the main manuscript, we present additional qualitative results for FIXMYPOSE (Fig. 1), PoseFixCPR (Fig. 2), and AIST-CPR (Fig. 3), respectively. The observed trend remains consistent: images that closely align with the provided reference image and transition description are ranked higher, while less relevant images are ranked lower.

Effectiveness of Stage I. To validate the effectiveness of Stage I in AutoComPose (i.e., body part-based descriptions), we modified the original prompts used by AutoComPose (Figures 9 and 10) to exclude body part-related queries (Figures 11 and 12). As shown in Table 1, removing Stage I results in a noticeable drop in performance, underscoring its critical role.

Smaller, more accessible MLLM. We conducted an experiment using GPT-4o mini (*gpt-4o-mini-2024-07-18*), a significantly smaller model that is over 16 times more cost-efficient than GPT-4o. As shown in Table 1, the impact of AutoComPose remains substantial, demonstrating its effectiveness and applicability even with lightweight models.

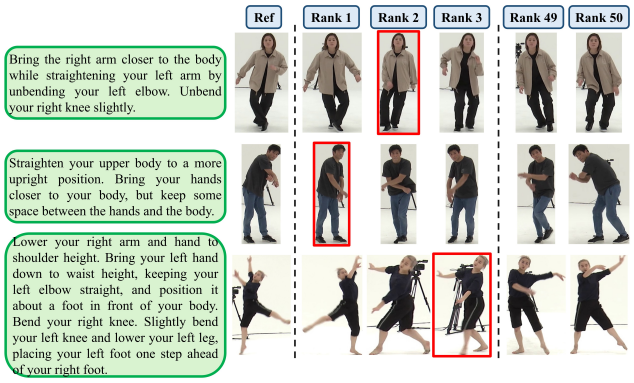

Figure 3. **Additional CPR results on AIST-CPR.** Ranked retrieval results are shown for a given reference image (Ref) and a transitional pose description (green box) as the query. Red boxes indicate the correct target images.

Pose Transition Descriptions	FIXMYPOSE (Size = 7106)			
	R@1	R@5	R@10	R@50
AutoComPose (full)	**8.24**	**27.45**	**38.63**	**63.53**
(-) Stage I	**8.63**	**26.67**	**37.84**	**62.75**
w/ GPT-4o mini	6.67	19.80	30.20	56.86
Human	0.20	1.57	2.94	13.53

Table 1. **Additional Ablation Study on AutoComPose.** The results are reported using the CLIP-RN50 setting.

2. Dataset Details

Triplet Examples. We showcase example pairs from the three datasets in Fig. 4.

Word Clouds. We compared word clouds generated from texts produced by AutoComPose with those from human annotators across the three datasets we used, as shown in Fig 5. Our findings indicate that texts generated by AutoComPose exhibit greater diversity and less bias.

FIXMYPOSE Filtering. A portion of the pose transition descriptions provided by FIXMYPOSE [5] contains environment-related instructions, such as guiding a pose transition by referring certain objects in a scene. Since our focus in this paper is on generating pose transition descriptions solely based on body movements, we utilized the multimodal large language model (MLLM) [4] with *Prompt-1* (Fig. 6) to detect and filter out such descriptions. Three examples of environment-related descriptions were included in the prompt to guide the filtering process.

PoseFixCPR Rendering. We constructed PoseFixCPR by rendering 2D images (including masks) obtained from 3D pose pairs from PoseFix [3] using Unreal Engine [1] based on the BEDLAM [2] rendering pipeline. A subset of assets used for rendering—including body meshes, body and clothing UV texture maps, and high-dynamic range panoramic images (HDRIs) that used for image-based

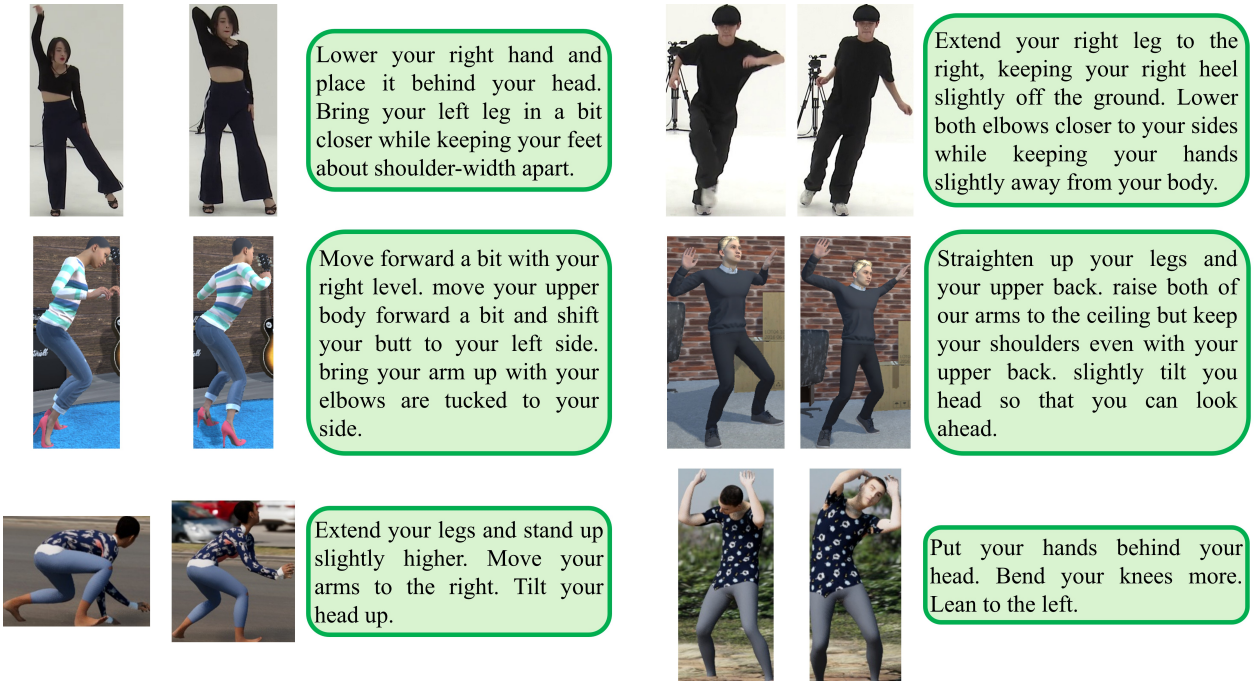

Figure 4. **Example pairs from three datasets.** The top, middle, and bottom rows correspond to the pairs from AIST-CPR, FIXMYPOSE, and PoseFixCPR, respectively.

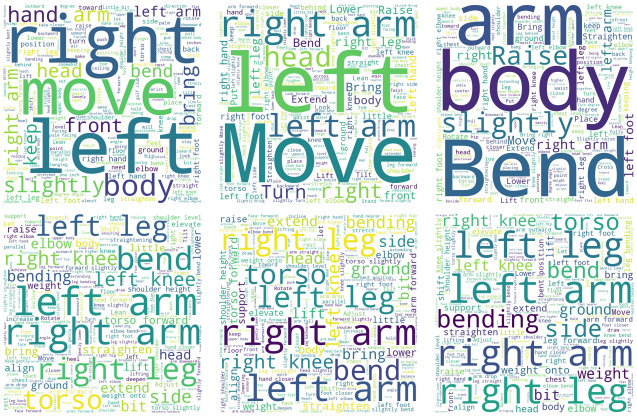

Figure 5. **Word clouds** generated from human-annotated texts (top) and AutoComPose-generated texts (bottom) across three datasets: FIXMYPOSE (left), PoseFixCPR (middle), and AIST-CPR (right).

lighting—is shown in Fig. 7. A selection of output images is displayed in Fig. 8. Note that only the human regions are used for composed pose retrieval (CPR), which can be extracted using the generated masks.

3. Generating Pose Transition Descriptions

MLLM Prompts. AutoComPose utilizes two prompts to communicate with the multimodal large language model (MLLM), as described in the main manuscript. *Prompt-2* (Fig. 9) is designed to guide the MLLM in generating body part-based pose transition descriptions. *Prompt-3* (Fig. 10) is used to helps the MLLM integrate these body part-based descriptions into more natural and complete sentences.

References

[1] Unreal Engine 5. https://www.unrealengine.com/en-US, 2025. 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 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1

[3] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 1

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[5] Hyounghun Kim, Abhay Zala, Graham Burri, and Mohit Bansal. Fixmypose: Pose correctional captioning and retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13161–13170, 2021. 1

Prompt-1

You will be given an instruction guiding an individual to transition from their current position to a target position. Your task is to determine whether the instruction includes any environmental direction descriptions—guiding the individual to move relative to a specific object nearby.
For example, the following instructions contain environmental direction descriptions, highlighted in quotation marks:
1. bring your right foot outwards at a 90 degree angle. bring your left and right hand up like a gun is being pointed at you ”keep your position facing the stereo system”.
2. ”lean your upper body forward to the plant pot” move the right leg to the right side a little bend the left leg a bit more move the right hand from waist level to beside the hip.
3. keep your feet and your hands on the floor as support push your arms up move your body and your legs up towards the ceiling ”lift up your head a little facing the rug”.

In contrast, the following instructions do not contain such descriptions:
1. position both your hands in front of your face is if you are grasping a pole. straighten the right leg so that the foot is flat on the floor. bend your left knee a little while pivoting on the ball of your foot. your head should face your hands.
2. first put your feet together take your right foot and put it slightly behind you. while keeping your lower body facing forward twist your upper body to the left. then hold your hands in the air at about the level of your head and turn your head to the left as well.
3. lift your right foot and pull it behind you. extend both of your hands gently outwards as if you are balancing on a surfboard while it is going under a big wave.

If the instruction contains environmental direction descriptions, respond with ”Yes.” Otherwise, respond with ”No.” Provide only the requested answer (”Yes” or ”No”) with no additional text before or after.

Figure 6. The prompt for removing environment-related descriptions in the FIXMYPOSE dataset.

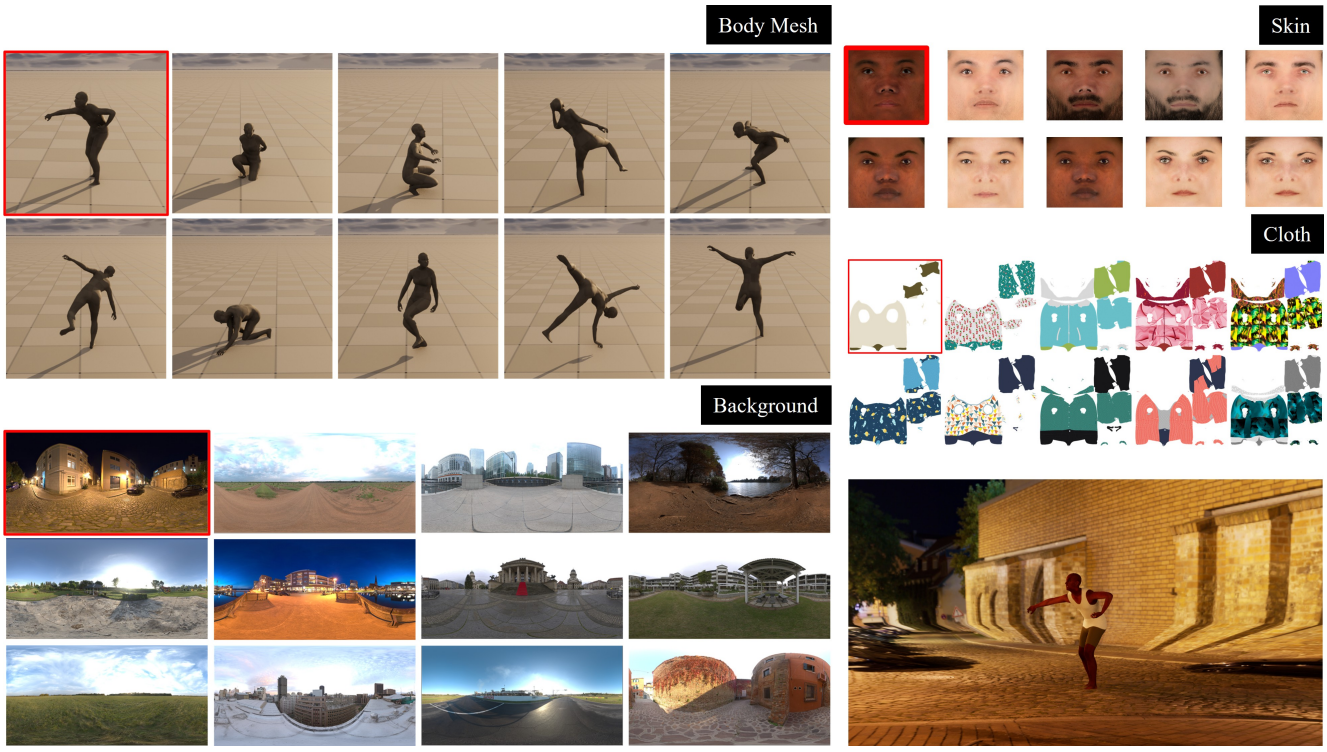

Figure 7. **Sampled assets used for constructing PoseFixCPR.** The highlighted assets (in the red box) were used to render the bottom-right image.

Figure 8. **Sampled images from PoseFixCPR.**

Prompt-2

You are a skilled fitness\dance instructor specializing in guiding individuals through pose transitions.
You will receive an image showing two poses side by side: the starting pose on the left and the target pose on the right.
Your task is to describe the transition from the starting pose to the target pose.
To achieve this:
- Compare the following body parts in the two poses: head, neck, left shoulder, right shoulder, left arm, right arm, left elbow, right elbow, left wrist, right wrist, left hand, right hand, torso, left hip, right hip, left leg, right leg, left knee, right knee, left ankle, right ankle, left foot, right foot.
- For each body part that requires adjustment, provide a clear and concise one-sentence description of the movement.
- Structure your description as follows: "1. Right Arm: Lift the right arm above the head, transitioning smoothly from its resting position." "2. Left Leg: Straighten the left leg fully as it supports the body's weight in standing position."

Below are some important guidelines that must be followed:
- Mention as few body parts as possible; omit subtle or nearly imperceptible movements.
- If the person is facing the camera, the left hand of the person is defined as the hand on your (the viewer's) right side.
- Provide only the requested descriptions of body part transitions, with no additional information before or after.

Figure 9. The prompt for generating body part-based pose transition descriptions.

Prompt-3

You will be given a set of bullet points describing a specific human pose transition.
Your task is to write five distinct, concise descriptions of the same pose transition. Structure your descriptions as follows: "Description 1: [description]." "Description 2: [description]."
Ensure your descriptions accurately capture the exact pose transition without altering it. To create variety, diversify your phrasing and avoid repeating expressions. You may incorporate analogies, such as "lower your right arm to your side as if you're holding a cane," as long as the described pose transition remains unchanged.

Figure 10. The prompt for integrating and diversifying body part-based pose transition descriptions.

Modified Prompt-2

You are a skilled fitness\dance instructor specializing in guiding individuals through pose transitions.
You will receive an image showing two poses side by side: the starting pose on the left and the target pose on the right.
Your task is to describe the transition from the starting pose to the target pose.
Provide a clear and concise description of the transition. Structure your description as follows: "Description: [description]."
Below are some important guidelines that must be followed:
- Omit subtle or nearly imperceptible movements.
- If the person is facing the camera, the left hand of the person is defined as the hand on your (the viewer's) right side.
- Provide only the requested descriptions of transitions, with no additional information before or after.

Figure 11. The modified Prompt-2 used in the ablation study for Stage I. The main modification, which removes body part-specific queries, is highlighted in red.

Modified Prompt-3

You will be given a description of a specific human pose transition.
Your task is to write five distinct, concise descriptions of the same pose transition. Structure your descriptions as follows: "Description 1: [description]." "Description 2: [description]."
Ensure your descriptions accurately capture the exact pose transition without altering it. To create variety, diversify your phrasing and avoid repeating expressions. You may incorporate analogies, such as "lower your right arm to your side as if you're holding a cane," as long as the described pose transition remains unchanged.

Figure 12. The modified Prompt-3 used in the ablation study for Stage I. The main modification, which removes body part-specific queries, is highlighted in red.