ControlPose: High-Fidelity Pose-Controlled Image Generation with Multi-Faceted Pose Disentanglement

Supplementary Material

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More method details in Sec. A.
- More dataset details in Sec. B.
- More evaluation metric details in Sec. C.
- More quantitative results in Sec. D.
- More qualitative results in Sec. E.
- Future work in Sec. F.

A. Method Details

A.1. Fourier Embedding

In the main paper, we state that the `Hierarchical Keypoint Layout Encoder` begins by processing raw keypoint coordinates. A standard Transformer architecture is permutation-invariant and does not inherently understand the continuous, spatial nature of 2D coordinates. To address this, we employ Fourier Embedding, a technique to transform low-dimensional, continuous coordinates into a higher-dimensional, periodic feature representation that is more suitable for a Transformer.

Given a single keypoint coordinate $\boldsymbol{p} = (x, y)$, where $x, y \in [0, 1]$ are normalized coordinates, we first define a set of d_f frequency bands. In our implementation, we set the number of bands $d_f = D_p/4$, where D_p is the target point embedding dimension. A frequency vector $\boldsymbol{\omega}$ is defined with a geometric progression of frequencies:

$$\boldsymbol{\omega} = [\omega_0, \omega_1, \ldots, \omega_{d_f-1}]^T, \tag{S1}$$

where the i-th frequency is $\omega_i = 100^{-i/d_f}$.

We apply these frequencies to both the x and y coordinates and compute their sine and cosine. This creates two feature vectors, $\gamma(x)$ and $\gamma(y)$, each of dimension $2d_f$:

$$\gamma(c) = [\sin(\omega_0 c), \cos(\omega_0 c), \ldots, \sin(\omega_{d_f-1} c), \cos(\omega_{d_f-1} c)]. \tag{S2}$$

The final embedding for the keypoint $\boldsymbol{p}$, which corresponds to the $\boldsymbol{e}_i$ in the main paper, is the concatenation of these two vectors:

$$\boldsymbol{e}_i = \text{FE}(\boldsymbol{p}) = \text{Concat}[\gamma(x), \gamma(y)]. \tag{S3}$$

The resulting feature vector $\boldsymbol{e}_i$ has a total dimension of $4d_f = D_p$. This high-dimensional representation allows the subsequent Transformer layers to more easily learn and attend to the relative spatial relationships between keypoints. For any non-existent keypoints, we use a dedicated learnable embedding $\boldsymbol{e}_{\text{null}}$ of the same dimension D_p.

A.2. Transformer Encoder

The `TransformerEncoder` mentioned in the main paper is responsible for aggregating the Fourier-embedded keypoint features into a single, representative vector for each predefined region. As shown in our code, this module is constructed from a stack of N_L standard Transformer encoder layers. Each layer ℓ consists of two main sub-layers: a Multi-Head Self-Attention (MHSA) module and a position-wise Feed-Forward Network (FFN).

Let the input sequence to layer ℓ be $\boldsymbol{Z}^{(\ell-1)} \in \mathbb{R}^{T \times D_p}$, where T is the sequence length (i.e., $1 + |\mathcal{R}_j|$). The computation flow for a single layer ℓ is as follows:

First, a residual connection and Layer Normalization are applied before the MHSA module:

$$\boldsymbol{Z}' = \text{MHSA}(\text{LayerNorm}(\boldsymbol{Z}^{(\ell-1)})) + \boldsymbol{Z}^{(\ell-1)}. \tag{S4}$$

The MHSA module itself operates by projecting the input into h attention heads, where h is the `nhead` parameter:

$$\text{MHSA}(\boldsymbol{Z}) = \text{Concat}[\text{head}_1, \ldots, \text{head}_h]W^O, \tag{S5}$$

$$\text{where head}_i = \text{Attention}(\boldsymbol{Z}W_i^Q, \boldsymbol{Z}W_i^K, \boldsymbol{Z}W_i^V). \tag{S6}$$

Second, the output of the attention module is passed through another residual connection, Layer Normalization, and the FFN:

$$\boldsymbol{Z}^{(\ell)} = \text{FFN}(\text{LayerNorm}(\boldsymbol{Z}')) + \boldsymbol{Z}'. \tag{S7}$$

The FFN, as defined in our code, is a two-layer perceptron with a GELU activation function. Its dimension is expanded by a factor of 4, consistent with standard Transformer implementations:

$$\text{FFN}(\boldsymbol{z}) = \text{GELU}(\boldsymbol{z}W_1 + b_1)W_2 + b_2, \tag{S8}$$

where $W_1 \in \mathbb{R}^{D_p \times 4D_p}$ and $W_2 \in \mathbb{R}^{4D_p \times D_p}$.

By stacking N_L such layers, the encoder iteratively refines the token representations. As noted in the main paper, we take the output of the '[CLS]' token from the final layer, $\boldsymbol{Z}_0^{(N_L)}$, as the rich, aggregated feature $\boldsymbol{h}_{\mathcal{R}_j,m}$ for that specific region.

A.3. Keypoint Grouping for Hierarchical Features

As mentioned in the main paper, the `Hierarchical Keypoint Layout Encoder` generates tokens at four

distinct levels: *Point-Level*, *Skeleton-Level*, *Half-Body-Level*, and *Full-Body-Level*. The *Point-Level* tokens correspond to the set of 17 keypoints (*e.g.*, 'nose', 'left_eye', 'left_shoulder', etc.). The higher levels are generated by aggregating features from predefined groups of these keypoints.

Rationale and Mechanism. The purpose of this hierarchical grouping is to provide the model with explicit, multi-scale semantic information about the human body. While raw keypoints provide precise location, they lack the intrinsic structural relationships that define a pose. For example, the model must *infer* that the 'left_shoulder', 'left_elbow', and 'left_wrist' keypoints form a "left arm." Our grouping strategy makes this relationship explicit. We use the Transformer Encoder mechanism described in Sec. A.2, where a `[CLS]` token is prepended to the features of a keypoint group (*e.g.*, $\boldsymbol{e}_{\text{left\_shoulder}}, \boldsymbol{e}_{\text{left\_elbow}}, \boldsymbol{e}_{\text{left\_wrist}}$). The corresponding output `[CLS]` token, $\boldsymbol{h}_{\mathcal{R}_j}$, thus serves as a single, aggregated feature vector representing the entire semantic concept (*e.g.*, the left arm).

Hierarchical Group Definitions. Our model uses the standard 17 keypoints from the COCO format, which are grouped into three higher-level hierarchies as follows:

- **Skeleton-Level:** This level partitions the body into its primary functional and anatomical components. It consists of 6 distinct regions:
  - `'head'`: ('nose', 'left_eye', 'right_eye', 'left_ear', 'right_ear')
  - `'torso'`: ('left_shoulder', 'right_shoulder', 'left_hip', 'right_hip')
  - `'left_arm'`: ('left_shoulder', 'left_elbow', 'left_wrist')
  - `'right_arm'`: ('right_shoulder', 'right_elbow', 'right_wrist')
  - `'left_leg'`: ('left_hip', 'left_knee', 'left_ankle')
  - `'right_leg'`: ('right_hip', 'right_knee', 'right_ankle')
- **Half-Body-Level:** This level provides the model with a clear concept of laterality by dividing the body vertically. It consists of 2 regions:
  - `'left_half'`: ('left_eye', 'left_ear', 'left_shoulder', 'left_elbow', 'left_wrist', 'left_hip', 'left_knee', 'left_ankle')
  - `'right_half'`: ('right_eye', 'right_ear', 'right_shoulder', 'right_elbow', 'right_wrist', 'right_hip', 'right_knee', 'right_ankle')
- **Full-Body-Level:** This level generates a single, holistic token that represents the entire person's pose, capturing the most global features.
  - `'full_body'`: (All 17 keypoints)

This process results in a total of $17 + 6 + 2 + 1 = 26$ distinct layout tokens for each person. As described in the main paper, learnable ID embeddings are added to this concatenated sequence before the final projection, allowing the model to distinguish between tokens representing individual points (*e.g.*, 'left_wrist') and those representing aggregated concepts (*e.g.*, 'left_arm').

B. Dataset Details

We validate our framework on three diverse, large-scale datasets that provide a comprehensive testbed for pose-guided human image generation, spanning artistic, real-world, and specialized domains.

Keypoint Definition. Across all datasets, we use the 17 keypoints defined by the COCO standard. This format provides a comprehensive representation of the human body, including facial features, limbs, and core body joints. The specific index and description for each keypoint are illustrated in Fig. S1.

Human-Art. The Human-Art dataset is a high-quality collection of 38,000 images, meticulously organized into 19 distinct scenarios. These scenarios are grouped into three primary domains, providing rich diversity in style and appearance. As illustrated in the directory structure, these domains include:

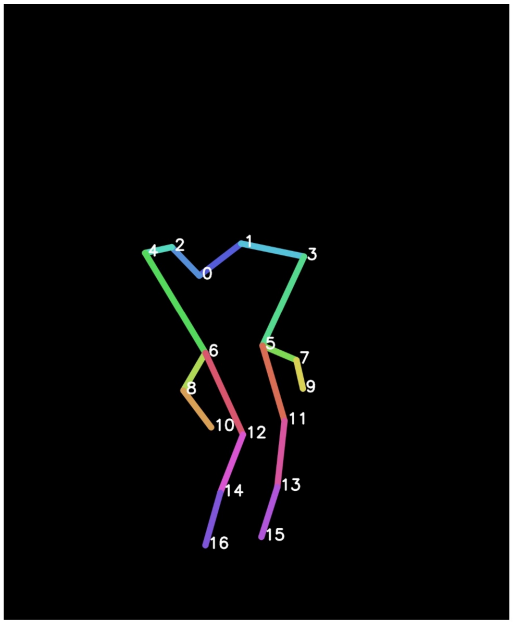
- `real_human`: Natural photographic images, including categories such as `acrobatics`, `cosplay`, `dance`, `drama`, and `movie` stills.
- `2D_virtual_human`: A wide array of 2D artistic styles, such as `cartoon`, `digital_art`, `ink_painting`, `kids_drawing`, `mural`, `oil_painting`, `shadow_play`, `sketch`, `stained_glass`, `ukiyoe`, and `watercolor`.
- `3D_virtual_human`: 3D-rendered imagery, including `garage_kits`, `relief`, and `sculpture`.

For our experiments, we adhere to the official train-validation split provided by the authors. The annotations are provided by the International Digital Economy Academy (IDEA) and are licensed under the Attribution-Non Commercial-Share Alike 4.0 International License (CC-BY-NC-SA 4.0).

LAION-Human. LAION-Human is a large-scale dataset derived from the LAION-5B database. It consists of approximately 1 million images that have been filtered based on human estimation confidence scores to ensure a high relevance to human subjects. Due to its massive scale, we utilize a curated subset for our training and validation. Specifically, we randomly sampled 200,000 images for our training set and 20,000 images for our validation set. This dataset is distributed under the Creative Common CC-BY 4.0 license, with individual image copyrights retained by their original creators.

UBC Fashion. The UBC Fashion dataset contains video sequences of fashion models performing full turns on a platform. To rigorously evaluate our model's ability to handle diverse, complex, and often infrequent poses, we process

(a) Sample image



(b) Sample pose

Keypoint	Description
0	none
1	left eye
2	right eye
3	left ear
4	right ear
5	left shoulder
6	right shoulder
7	left elbow
8	right elbow
9	left wrist
10	right wrist
11	left hip
12	right hip
13	left knee
14	right knee
15	left ankle
16	right ankle

(c) Definition of keypoints

Figure S1. **Illustration of the 17-point COCO pose format used in our work. (a)** A sample image from the Human-Art dataset. **(b)** The corresponding pose skeleton extracted from the image. Each distinct body part (e.g., left arm, right leg) is marked with a different color for visual clarity. **(c)** The explicit definition and index mapping for the 17 keypoints used throughout our experiments, following the standard COCO format.

these sequences to extract individual frames. We specifically sample frames representing three distinct orientations: `front`, `side`, and `back`. This process yields a specialized test set of approximately 1,200 frames for each of the three orientations. The dataset is available under the Creative Commons Attribution-Non Commercial 4.0 International Public License.

C. Evaluation Metric Details

We evaluate our model's performance using a comprehensive suite of metrics designed to measure pose fidelity, perceptual quality, and semantic alignment, as outlined in our main experimental tables.

Pose Accuracy. This category of metrics measures how faithfully the generated image adheres to the input pose conditions, both in terms of keypoint location and the number of subjects.

- **Average Precision (AP).** This is our primary metric for pose fidelity. We first use a pre-trained pose estimator to detect the 17 COCO keypoints from each generated image. We then evaluate the accuracy of these detected keypoints against the ground truth input keypoints using the standard COCO-style Average Precision (AP) metric. This metric relies on Object Keypoint Similarity (OKS) to determine a match. OKS measures the normalized Euclidean distance between a detected keypoint $\hat{\boldsymbol{p}}_i$ and its corresponding ground truth $\boldsymbol{p}_i$:

$$\text{OKS} = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2) \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \tag{S9}$$

where $d_i = \|\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i\|_2$ is the distance, s is the object scale, k_i is a keypoint-specific constant, and v_i is the visibility flag of the ground truth keypoint. We report the standard mean AP, averaged over OKS thresholds from 0.50 to 0.95 (denoted as AP).

- **Cosine Similarity-based AP (CAP).** While OKS-based AP measures spatial displacement, we also evaluate the fidelity of the pose's overall structure, which is less sensitive to translation or scaling errors. For this, we compute a normalized feature vector for both the ground truth and detected poses. First, all visible keypoint coordinates (X and Y sets treated separately) are normalized to a $[0, 1]$ range and then L2-normalized. These normalized coordinate sets are concatenated into a single feature vector ($\boldsymbol{v}_{\text{gt}}$ for ground truth, $\boldsymbol{v}_{\text{det}}$ for detection). The similarity score is the cosine similarity between these two vectors:

$$\text{Sim}(\boldsymbol{v}_{\text{gt}}, \boldsymbol{v}_{\text{det}}) = \frac{\boldsymbol{v}_{\text{gt}} \cdot \boldsymbol{v}_{\text{det}}}{\|\boldsymbol{v}_{\text{gt}}\| \|\boldsymbol{v}_{\text{det}}\|}. \tag{S10}$$

The CAP metric is an AP variant that uses this similarity score as the matching threshold, with the main reported metric being the average over similarity thresholds from 0.50 to 0.95.

Table S1. Detailed results on UBC Fashion dataset with front, side, and back orientations.

Orientation	Method	Pose Accuracy			Image Quality		T2I Alignment
		AP(%) ↑	CAP(%) ↑	PCE ↓	FID ↓	KID ↓	CLIP-Score ↑
Front	T2I-Adapter [23]	71.34	51.41	0.63	6.82	4.08	32.17
	ControlNet [37]	74.92	47.89	0.64	**4.71**	3.98	32.10
	Uni-ControlNet [39]	80.15	51.36	0.58	5.95	4.33	32.19
	HumanSD [13]	77.12	51.72	0.62	6.43	4.97	31.14
	Stable-Pose [33]	88.69	51.68	**0.57**	5.89	4.41	32.05
	GRPose [35]	88.12	51.78	0.62	6.11	4.32	32.31
	SP-Ctrl [34]	87.45	51.67	0.74	6.01	3.99	32.18
	EasyControl [38]	89.12	52.93	0.79	6.08	3.97	32.22
	OminiControl [30]	88.93	52.89	0.72	6.11	4.03	31.89
	ControlPose(ours)	**89.78**	**53.11**	0.59	5.72	**3.94**	**32.45**
Side	T2I-Adapter [23]	37.15	44.52	0.74	5.98	**3.76**	32.16
	ControlNet [37]	53.04	46.29	0.78	5.17	4.02	32.13
	Uni-ControlNet [39]	58.51	46.77	**0.70**	5.62	4.19	32.15
	HumanSD [13]	58.45	45.61	0.76	5.23	4.68	31.95
	Stable-Pose [33]	69.83	47.03	0.73	5.08	4.51	**32.27**
	GRPose [35]	69.89	46.81	0.78	5.12	4.45	32.23
	SP-Ctrl [34]	68.27	46.73	0.74	5.33	4.48	32.14
	EasyControl [38]	69.34	46.97	0.77	4.97	4.14	32.18
	OminiControl [30]	69.14	47.01	0.73	**4.92**	4.09	32.20
	ControlPose(ours)	**69.97**	**47.05**	0.72	5.07	3.99	32.23
Back	T2I-Adapter [23]	5.89	22.51	0.93	8.49	4.34	32.22
	ControlNet [37]	24.12	24.78	0.95	7.53	4.09	32.15
	Uni-ControlNet [39]	20.24	23.61	0.91	7.20	4.28	32.19
	HumanSD [13]	10.27	21.37	0.89	9.12	4.99	32.01
	Stable-Pose [33]	30.14	25.35	0.94	6.24	4.56	32.11
	GRPose [35]	28.58	24.18	0.98	6.21	4.21	32.07
	SP-Ctrl [34]	29.34	25.73	0.94	6.30	4.18	32.11
	EasyControl [38]	29.11	26.11	0.90	6.18	4.05	32.14
	OminiControl [30]	30.32	26.08	**0.88**	**6.11**	**3.99**	32.06
	ControlPose(ours)	**61.54**	**30.12**	0.95	6.16	4.02	**32.23**

- **People Counting Error (PCE).** Given our focus on multi-person synthesis, we evaluate the model's ability to generate the correct number of subjects. We employ a pre-trained human detector on the generated images. PCE is the Mean Absolute Error between the number of detected persons N_{det} and the ground truth number of input poses N_{gt}:

$$\text{PCE} = \mathbb{E}[|N_{\text{det}} - N_{\text{gt}}|]. \tag{S11}$$

Image Quality. To assess the realism and visual fidelity of generated images, we use two standard distribution-based metrics. The real images are sourced from the validation set (*e.g.*, Human-Art).

- **Fréchet Inception Distance (FID).** FID measures the dissimilarity between the distribution of real images ($\boldsymbol{X}$) and generated images ($\boldsymbol{G}$). It models the features extracted from an Inception v3 model as multivariate Gaussian distributions. The distance is computed as:

$$\begin{aligned}\text{FID}(\boldsymbol{X}, \boldsymbol{G}) = &\|\boldsymbol{\mu_x} - \boldsymbol{\mu_g}\|_2^2 \\ &+ \text{Tr}\left(\boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_g} - 2(\boldsymbol{\Sigma_x}\boldsymbol{\Sigma_g})^{1/2}\right),\end{aligned} \tag{S12}$$

where $(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})$ and $(\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})$ are the mean and covariance of the Inception v3 features (from the 64-dim layer) for the real and generated distributions, respectively. A lower FID score indicates a smaller distance between the two distributions.

- **Kernel Inception Distance (KID).** As FID can be sensitive to small sample sizes, we also report the Kernel Inception Distance (KID). KID computes the squared Maximum Mean Discrepancy (MMD) between the feature distributions. Using a third-degree polynomial kernel $k(\boldsymbol{a}, \boldsymbol{b}) = (\frac{1}{d}\boldsymbol{a}^T\boldsymbol{b} + 1)^3$, the squared MMD is computed

as:

$$\text{MMD}^2 = \mathbb{E}_{\boldsymbol{a},\boldsymbol{a}'\sim\mathcal{F}_{\boldsymbol{x}}}[k(\boldsymbol{a},\boldsymbol{a}')] + \mathbb{E}_{\boldsymbol{b},\boldsymbol{b}'\sim\mathcal{F}_{\boldsymbol{g}}}[k(\boldsymbol{b},\boldsymbol{b}')] - 2\mathbb{E}_{\boldsymbol{a}\sim\mathcal{F}_{\boldsymbol{x}},\boldsymbol{b}\sim\mathcal{F}_{\boldsymbol{g}}}[k(\boldsymbol{a},\boldsymbol{b})]. \quad \text{(S13)}$$

Here, $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{F}_{\boldsymbol{g}}$ are the sets of Inception features from real and generated images, respectively. $\boldsymbol{a}$ and $\boldsymbol{a}'$ represent independent samples drawn from the real distribution $\mathcal{F}_{\boldsymbol{x}}$, while $\boldsymbol{b}$ and $\boldsymbol{b}'$ are independent samples drawn from the generated distribution $\mathcal{F}_{\boldsymbol{g}}$.

T2I Alignment. This metric evaluates the semantic alignment between the output image and the input text prompt.

- **CLIP Score.** To measure how well the generated image $\boldsymbol{g}$ aligns with the input text prompt $\boldsymbol{y}$, we compute the CLIP Score. We use the pre-trained CLIP-ViT-B/16 model, which provides a text encoder τ_T and an image encoder τ_I. The score is the scaled cosine similarity between the two feature embeddings:

$$\mathcal{S}_{\text{CLIP}} = 100 \cdot \cos(\tau_I(\boldsymbol{g}), \tau_T(\boldsymbol{y})). \quad \text{(S14)}$$

D. Quantitative Results

We provide a detailed breakdown of performance on the UBC Fashion dataset in Tab. S1. For improved readability, all reported KID scores have been scaled by 100. The results clearly show that our ControlPose consistently achieves superior controllability across all orientations, including challenging and infrequently encountered poses.

E. Qualitative Results

We showcase more extensive qualitative results from our method to demonstrate its robustness and versatility. Figure S2 presents additional results on the diverse HumanArt dataset, while Figure S3 shows more examples on the LAION-Human dataset.

These results collectively illustrate that ControlPose consistently generates high-fidelity images with strong pose alignment, even under challenging conditions that are often problematic for other methods. Our model successfully handles complex multi-person interactions, intricate actions, and non-standard body orientations, maintaining clear separation between subjects and accurately rendering the specified poses.

F. Future Work

While ControlPose demonstrates significant advancements in instance-aware 2D image synthesis, its core principles of feature disentanglement and structured control offer promising avenues for several high-dimensional and more complex domains. A primary future direction is to extend our framework from static images to video generation. The key challenge in video synthesis is maintaining temporal consistency, not only in the background but also in the appearance and motion of each subject. Our instance-aware architecture provides a strong foundation for this task. The Individual Pose Encoder could be evolved to manage a subject's identity and appearance features across frames, while the Hierarchical Layout Encoder could be adapted to process spatio-temporal trajectories, ensuring smooth and coherent motion. Another exciting avenue is the application of our principles to 3D-aware synthesis. Instead of 2D keypoints, the Hierarchical Layout Encoder could be re-engineered to process 3D pose representations, such as SMPL parameters or 3D joint coordinates. By integrating this 3D-native pose control into generative models like generative Neural Radiance Fields (NeRFs), our method could enable the creation and manipulation of 3D human assets or the rendering of 2D images from novel camera angles, all guided by precise 3D poses.

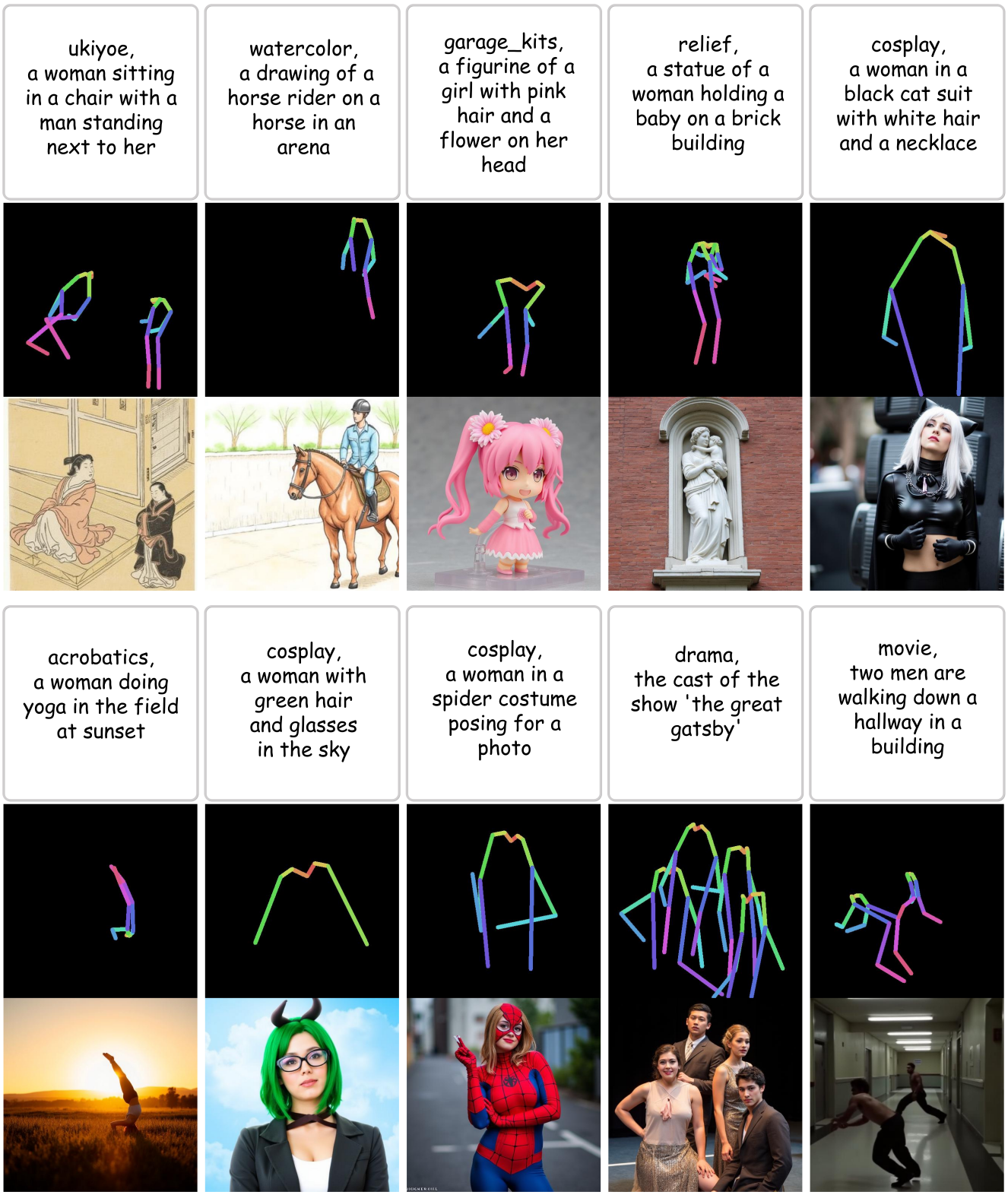


Figure S2. **More qualitative results on Human-Art datasets.**

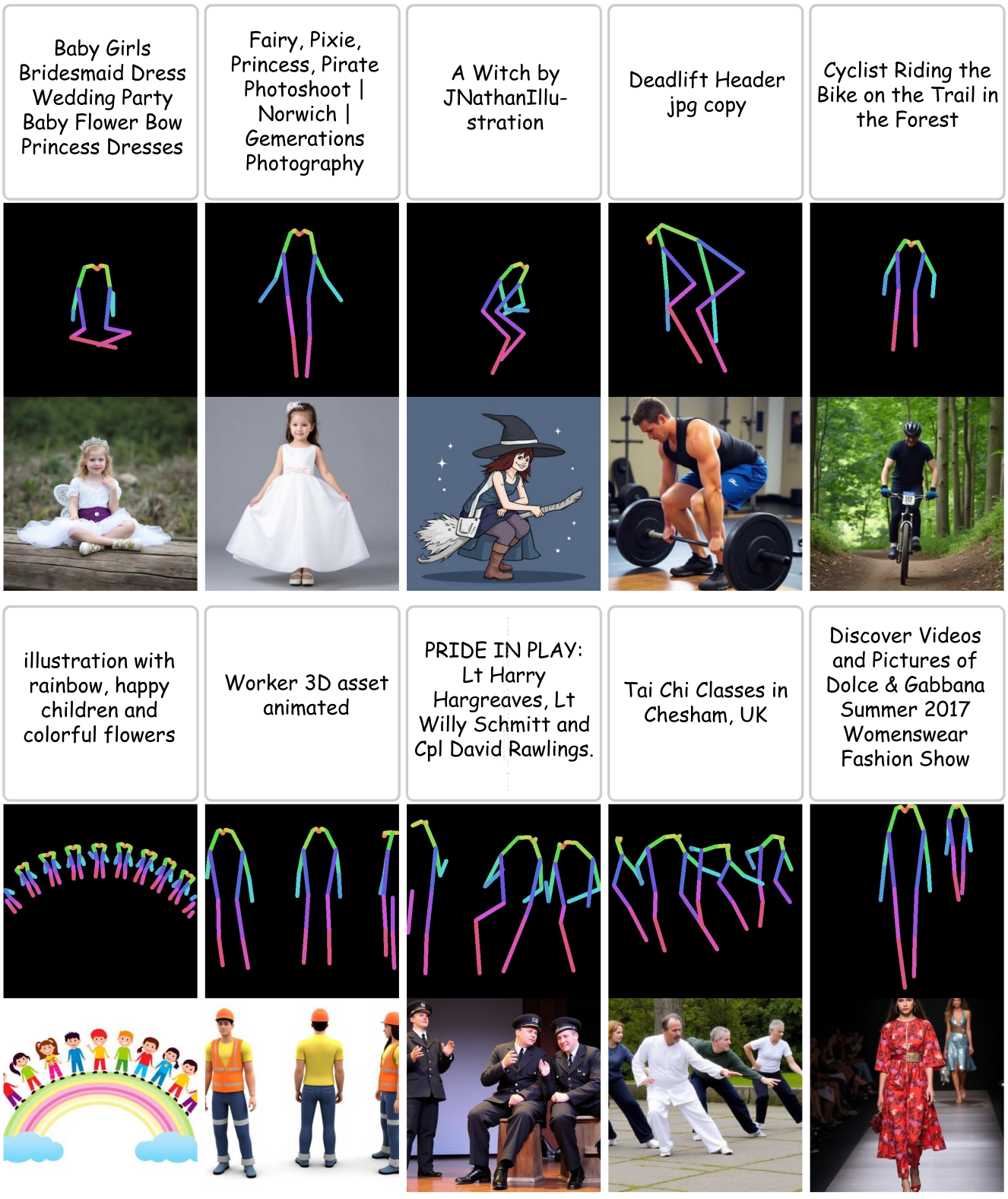


Figure S3. **More qualitative results on LAION-Human datasets.**