Supplementary Material of
FineControlNet: Fine-level Text Control for Image Generation with Spatially Aligned Text Control Injection

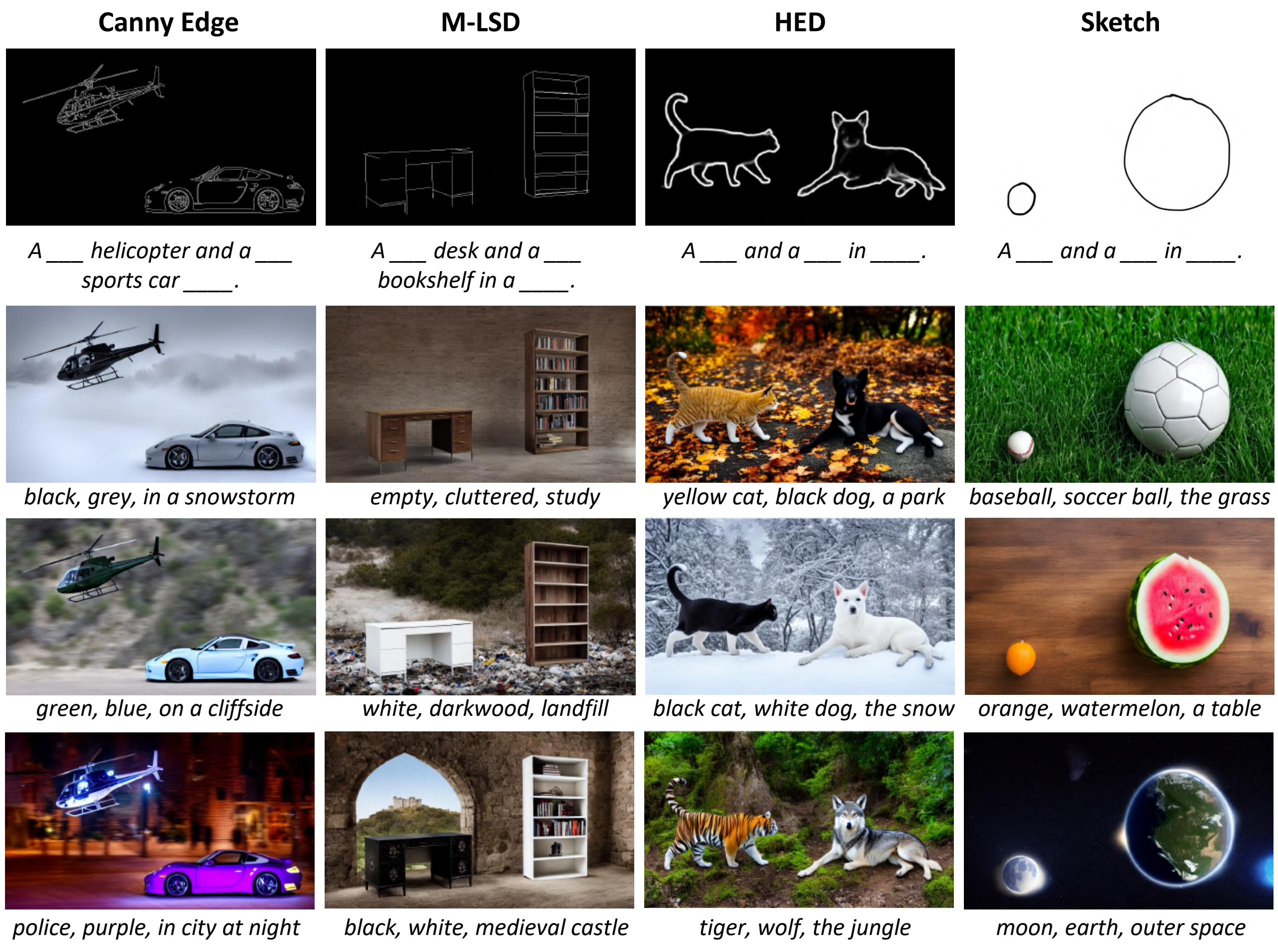


Figure 1. Results of FineControlNet applied to different control modalities of Canny [2] edges, M-LSD [6] lines, HED [13] edges, and a sketch input. As shown above our method has the ability to not only work on human pose inputs, but other modalities as well using the same approach described in our method section but applied to different ControlNet [14] models. Under each column is the modality name, the sample input image, the prompt template, and three examples images with the corresponding input prompt information. Our method demonstrates the ability to finely control each instance.

In this supplementary material, we showcase additional experimental results that could not be included in the main paper due to space constraints.

1) We demonstrate our framework's versatility by extending it to different control modalities. 2) We elaborate on the two stages of our prompt parsing approach: "Image Context Separation Task" and "Match Instance to Coordinate." 3) We provide statistics of our newly created dataset. The dataset file in JSON format is included in our ZIP archive. Subsequently, 4) We further provide the details of our pose control evaluation to prevent confusion and discuss its limitations. 5) We present the results of a ControlNet [14] variant, Multi-ControlNet, and analyze its limitations in comparison to our method. 6) We provide an extensive qualitative comparison with ControlNet, our direct baseline. 7) We conduct a robustness study of our method

against different factors, which will aid in its practical application. 8) We further compare FineControlNet with Flux [9] and SDXL [1]. 9) We discuss the limitations of our method and outline potential future work. 10) We reflect on the societal impact of our work and discuss our ethical considerations.

1. Different Control Modality

We present results demonstrating the efficacy of our FineControlNet architecture using various geometric control modalities, including Canny [2] edges, M-LSD [6] lines, HED [13] edges, and sketch inputs. As illustrated in Figure 1, our framework enables fine-grained text-based control over individual instances while maintaining coherence across the generated scene. Through spatially aligned text injection, each instance faithfully reflects the corresponding textual prompt, with harmonized style and lighting that is consistent both within and between instances. For example, the bottom left image generated from the prompt "A police helicopter and a purple sports car in city at night" supports these claims; both vehicles exhibit glossy textures and lighting congruent with the nocturnal urban setting.

2. Prompt Parsing

Here we include our full prompts for parsing the global description and assigning it to each skeleton, as well as a typical output from ChatGPT 4.0 responding to the prompts.

1) Below is a real exchange of prompts between our method and ChatGPT 4.0 for the purpose of parsing a global prompt into individual descriptions. The user's comments are highlighted in blue and ChatGPT's comments are highlighted in green.

User:
Please help me with a task.
During this task, do not respond with anything other than what is instructed of you.
Any response from you will be regarded as your output to the input.
Here is the task:
Given an input string of text that describes an image with multiple people in it along with some background information, separate the global context into individual descriptions WITHOUT spatial prepositions.

For example, an input might look like:
"Two people on a rooftop, a construction worker on the left and a man in a suit on the right"

And your output should look like:
Person 1 Description: "A construction worker on a rooftop"
Person 2 Description: "A man in a suit on a rooftop"

One more example:
Input: "A cowboy on the left shaking hands with a doctor on the right in times square"

Output:
Person 1 Description: "A cowboy shaking hands in times square"
Person 2 Description: "A doctor shaking hands in times square"
If you understand respond with "I understand" and the next message will be the first input.

ChatGPT:
I understand

User:
Spider-man swingin in from the left with green goblin on the far right. A policeman is between them trying to stop the fight in the street.

ChatGPT:
Person 1 Description: "Spider-man swinging in the street"
Person 2 Description: "Green Goblin in the street"
Person 3 Description: "A policeman trying to stop the fight in the street"

2) We also include a real example of using ChatGPT to assign the parsed individual prompts to the center-points of skeletons represented as 2D coordinates. The initial global description is also provided in order to give context as to which individual description corresponds to which coordinate point. Again, the user's comments are highlighted in blue and ChatGPT's comments are highlighted in green.

User:
Please help me with a task.
Do not make or run any python code. Only use language to describe your thinking and then output the final response.
Here is the task:
Given a global description, a list of instance descriptions, and a list of x,y coordinates of the same length, return a dictionary that matches one instance description to one x,y coordinate.

Example 1:
Input:
global_desc = ”Two people on a rooftop, a construction worker on the left and a man in a suit on the right”
instance_desc = [”A construction worker on a rooftop”, ”A man in a suit on a rooftop”]
coordinates = [[100, 50], [300, 41]]

match_instance_to_coordinate(global_desc, instance_desc, coordinates)
Output:
”A construction worker on a rooftop”: [100, 50], ”A man in a suit on a rooftop”: [300, 41]

Example 2:
Input:
global_desc = ”3 children in a park. The one on the left is wearing a blue shirt, the one in the middle is wearing a red shirt, and the one on the right is wearing a green shirt.”
instance_desc = [”A child in a blue shirt in a park”, ”A child in a red shirt in a park”, ”A child in a green shirt in a park”]
coordinates = [[75, 80], [28, 60], [344, 75]]

match_instance_to_coordinate(global_desc, instance_desc, coordinates)
Output:
”A child in a blue shirt in a park”: [28, 60], ”A child in a red shirt in a park”: [75, 80], ”A child in a green shirt in a park”: [344, 75]

If no specific location is given for each person, you may assign them randomly. If the number of people does not match the number of coordinates, output an error.
If you understand respond with ”I understand” and the next message will be the first input.

ChatGPT:
I understand

User: global_desc = ”Spider-man swingin in from the left with green goblin on the far right. A policeman is between them trying to stop the fight in the street.”
instance_desc = [”Spider-man swinging in the street”, ”Green Goblin in the street”, ”A policeman trying to stop the fight in the street”]
coordinates = [[92, 74], [15, 140], [57, 77]]

ChatGPT: To match each instance description to one of the given coordinates, we need to interpret the spatial arrangement described in the global description and then allocate the coordinates accordingly.

From the global description, we have three key figures: Spider-man, Green Goblin, and a policeman. Their positions relative to each other are as follows: Spider-man is coming in from the left, Green Goblin is on the far right, and the policeman is between them.

Now, let’s consider the coordinates. They are given as [[92, 74], [15, 140], [57, 77]]. Without specific indications of what these coordinates represent (like left-to-right or up-to-down orientation), we’ll assume a standard coordinate system where the first number (x-coordinate) increases as we move to the right, and the second number (y-coordinate) could represent vertical positioning.

Based on this assumption:
Spider-man, coming in from the left, is likely to have the smallest x-coordinate. Green Goblin, on the far right, is likely to have the largest x-coordinate. The policeman, being in between, should have an x-coordinate that is between the other two.

Looking at the provided coordinates:
[15, 140] has the smallest x-coordinate. [92, 74] has the largest x-coordinate. [57, 77] falls in between these two.

Thus, the matching would be:
Spider-man (coming in from the left): [15, 140] Green Goblin (on the far right): [92, 74] Policeman (in between): [57, 77]

The final output dictionary would be:

{ ”Spider-man swinging in the street”: [15, 140], ”Green Goblin in the street”: [92, 74], ”A policeman trying to stop the fight in the street”: [57, 77] }

We tested ChatGPT-4’s ability to both parse and assign individual prompts from the global prompt on our dataset. For parsing global prompts to instance descriptions, we found ChatGPT-4 was capable of correctly parsing 1121 / 1126 or 99% of the prompts in our dataset. With the main failure case being sometimes appending the scene information to the individual descriptions, ie. “a doctor in the desert” instead of just “a doctor”.

For the task of matching instance descriptions to skeleton locations in the image we also used ChatGPT-4 on our dataset, and found that it was capable of correctly matching

1084 / 1126 or 96% of sets of instance descriptions to the corresponding skeletons correctly. The main failure case for this task was ChatGPT-4 incorrectly formatting it's response, leaving it unusable for downstream assigning of the prompt to the skeleton. For example a good output would end with a list ordering the descriptions from left to right, ie. "... ['Darth Vader', 'Spider-Man']", while a bad output may simply describe the order of descriptions, ie. "... this results in Spider-Man on the right and Darth Vader on the left".

3. Dataset

We provide the histograms of numbers of people per image, person's bounding box resolution per image area ratio, and CrowdIndex [10] in Figure 2, for our dataset. CrowdIndex computes the ratio of the number of other persons' joints against the number of each person's joints. Higher CrowdIndex indicates higher chance of occlusion and interaction between people. The low resolution ratio and the higher CrowdIndex are related to the difficulty of identity and pose control due to discretization in latent space and ambiguity of instance assignment in attention masks.

4. Pose Control Evaluation

To evaluate the pose control accuracy of methods, we test HigherHRNet [3] on generated images following HumanSD [8]. HigherHRNet is the state-of-the-art multi-human 2D pose estimator, and the weights are trained on MSCOCO [11] and Human-Art [7] by the authors of HumanSD. We report the average precision (AP) of Object Keypoint Similarity (OKS) [11] measured in different distance thresholds. The superscript categorizes the resolution of people in an image and measures the average precision only for persons in that category, where M and L denote *medium* and *large*, respectively.

We also evaluate people count accuracy using two metrics: People Count Error (PCE) and Human Number Difference (HND), following [4, 8] using the HigherHRNet. PCE measures the average binary error in people count per generated image, with 1 indicating an incorrect count and 0 a correct count. HND reports the average difference between the number of ground truth input skeletons and the number of detected skeletons in the generated image. Both PCE and HND effectively capture false positive rates when generating images containing multiple people. Note that AP, PCE, and HND are pseudo metrics, because they are susceptible to inaccuracies of the 2D pose estimator independent from the inaccuracy of image generation. Last, the PCE metric reported in the paper of Ju et al. [8] is different from the original PCE [4] metric and equals to HND metric in this manuscript.

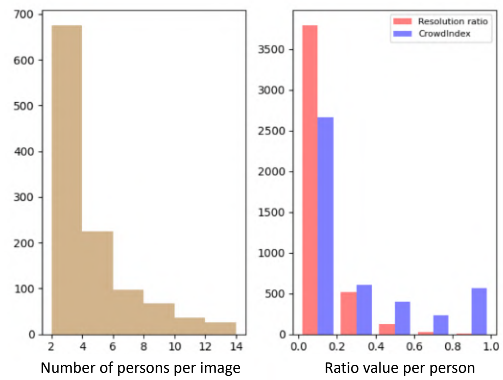


Figure 2. Statistics of our dataset. The y-axis indicates the counts that fall in bins in the x-axis.

5. Difference with MultiControlNet

We compare FineControlNet to MultiControlNet [14], an extension of ControlNet supporting multiple geometric modalities (e.g. pose, depth) with a single text prompt. For equivalence, we modify MultiControlNet to condition on instance-specific texts over multiple poses. Experiments utilize a third-party HuggingFace Diffusers [5] implementation. Results in Figure 3 demonstrate compromised adherence to per-instance textual prompts compared to FineControlNet, stemming from lack of spatial text-latent alignment and careful latent composition. Moreover, MultiControlNet fails to process more than two inputs, generating blurry and abstract imagery. These contrasts highlight the importance of FineControlNet's spatially aware text injection and carefully engineered latent fusion for fine-grained, multi-instance control.

6. More Qualitative Results

Additional qualitative results of FineControlNet's ability to address instance-specific constraints are shown in Figures 6 and 7. The input poses and prompts are shown in the leftmost columns and at the bottom of each row of images, respectively. The results of FineControlNet are provided in the middle two columns, with and without the poses overlaid on the generated images. We also show the outputs of ControlNet [14] using the same pair of input poses and text prompts as a reference for comparison in the rightmost columns. For both methods, we use the same seed numbers which are sampled from a uniform distribution.

7. How Robust is FineControlNet?

We analyze the robustness of FineControlNet to variations in number of people, scale, and inter-personal distance. Quantitative experiments recording CLIP Identity Observance (CIO) scores (Table 1) and qualitative results

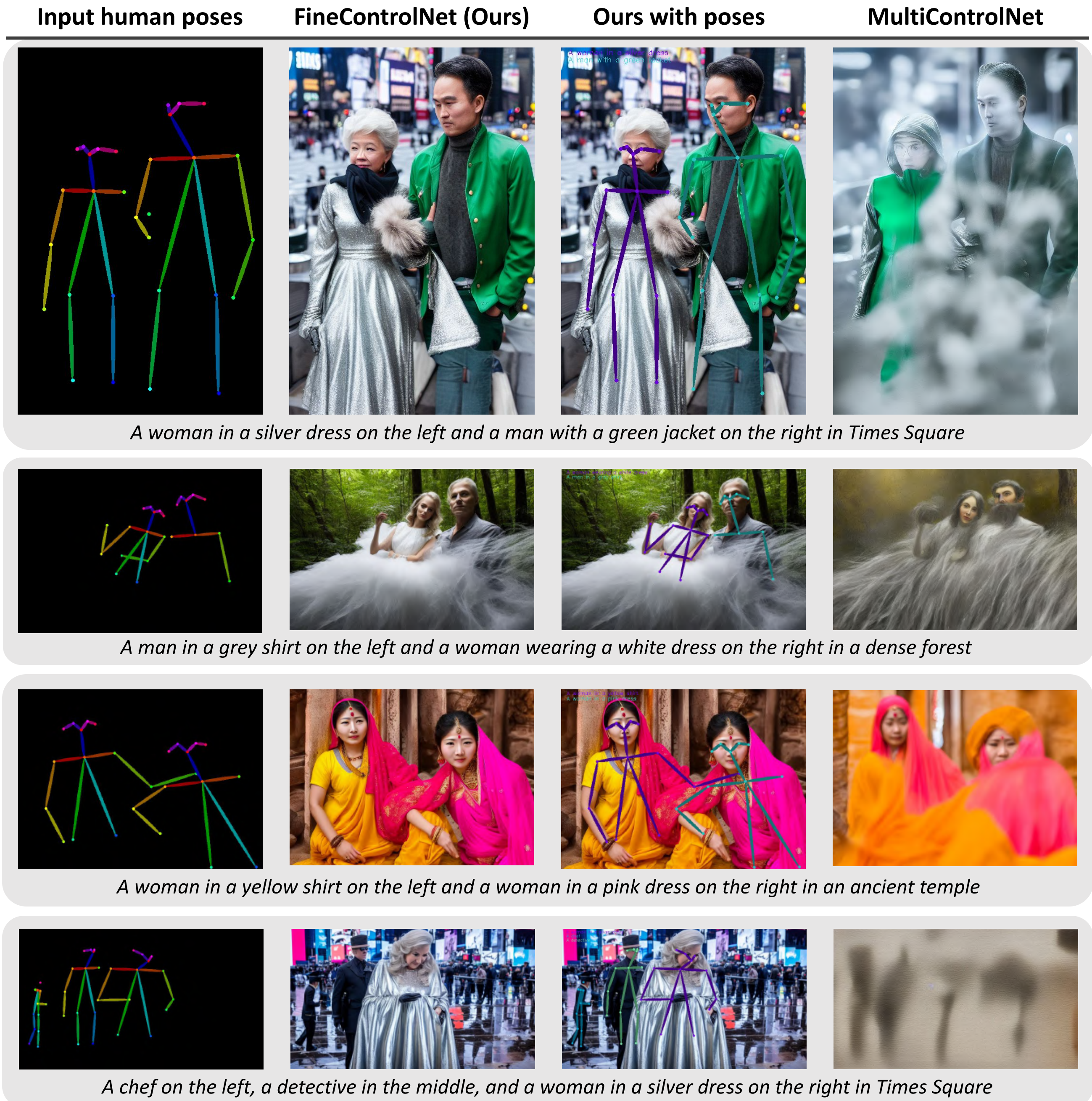


Figure 3. Comparison between our FineControlNet and MultiControlNet [5, 14]. MultiControlNet produces blurry images, which also have blended appearance/identity between instances. In addition, more than two geometric control inputs paired with different text prompts often cause a complete failure. We provide the images of poses overlaid on FineControlNet's generated outputs for reference.

(Figures 8-10) demonstrate performance under differing conditions.

Varying the number of input 2D poses while fixing scale and spacing reveals strong text-image consistency for 2-3 people, with gradual degradation as count increases to 5 and 7 (Table 1; Figure 8). For instance, the fourth person from the left in Figure 8 fails to wear the prompted dress, illustrating compromised identity observance. We posit that as instance count rises, pressure to balance identity adherence against holistic visual harmonization becomes more severe,

Figure 4. Comparison with “SDXL with Prompt Weighting without ControlNet” and “Flux without Prompt Weighting with ControlNet”. The leftmost column shows input (a text prompt, 2D poses, and an original image where 2D poses are extracted). The first row results are from SDXL. The second row results are from Flux. The third row results are from FineControlNet. Despite FineControlNet utilizing a comparatively less advanced backbone model for generation, it demonstrates superior ability to preserve the identity of given prompts, outperforming state-of-the-art techniques in terms of instance-level control.

Figure 5. Further comparison with “Flux with ControlNet.” The leftmost column shows the 2D pose input and the source image. The text prompt is “From left to right: a metal humanoid, Darth Vader, and a medieval knight in the desert.” The first row presents results from Flux, while the second row shows results from FineControlNet. The comparison reveals that while Flux demonstrates a reasonable ability to preserve identity from text prompts, it exhibits difficulties in consistently associating instance descriptions to corresponding pose inputs (the second and third outputs of Flux). In addition, it occasionally blends visual features (the first and third outputs of Flux).

Table 1. Robustness Study regarding factors of "number of people", "scale of a person", and "distance between people".

Metrics	Number of People			Scale of a Person					Distance between People			
	3	5	7	1	0.75	0.5	0.25	0.1	1	0.75	0.5	0.25
$\text{CIO}_{\text{sim}} \uparrow$	28.2	26.9	26.5	28.2	27.5	26.4	23.2	20.3	28.2	27.8	27.8	25.3
CIO_{σ}	0.74	0.46	0.32	0.74	0.69	0.62	0.55	0.42	0.74	0.7	0.69	0.48
$\text{CIO}_{\text{diff}} \uparrow$	5.3±2.4	3.2±1.9	2.2±1.3	5.3±2.4	4.6±2.5	3.6±1.9	2.0±1.4	0.9±0.7	5.3±2.4	4.8±2.3	4.6±2.6	2.2±1.3

increasing feature sharing between instances.

Experiments assessing robustness to variations in human scale utilize three input poses while fixing inter-personal distances. As depicted in Figure 9 and Table 1, identity observance degrades gradually with increased downscaling, tied to losses in latent feature resolution. Performance remains reasonable down to 50% scale, with more significant drops emerging under extreme miniaturization. Note input pose map resolution is constant at 512 pixels in height.

Similarly, distance experiments alter spacing around a central pose with three total people at fixed scale. Results in Figure 10 and Table 1 demonstrate consistent identity retention given non-overlapping inputs, with overlap introducing instance dropping or blending.

Together, these analyses quantify trade-offs between fidelity and spatial configurations. Performance gracefully handles reasonable perturbations but breaks down at data distribution extremes. Addressing such generalization limits highlights an area for further improvement. Note that the evaluation data we used for the robustness study is different from the new dataset we introduced in our main text. We validated each condition on 100 data points, which use different random seeds per data point for generation with other hyperparameters fixed.

8. Further Comparison with recent methods

The recent techniques of prompt weighting and Flux [9] have been shown to be effective in generating images that accurately adhere to text prompts. We evaluated FineControlNet in comparison with "SDXL [1] with Prompt Weighting without ControlNet" and "Flux without Prompt Weighting with ControlNet." For the implementation, we employed the Diffusers API[1] for SDXL, Flux, and ControlNet, along with the Compel API[2] for prompt weighting, which is widely adopted within the community. However, it is important to note that Compel does not officially support either ControlNet or Flux. Additionally, our FineControlNet utilizes the smallest and oldest backbone model, resulting in comparatively lower perception quality.

As shown in Figure 4, SDXL with prompt weighting fails to adequately maintain the distinct identities specified in the prompts. In some instances, a single identity dominates the entire image, being applied uniformly across all subjects, while in other cases, there is noticeable blending, such as the merging of a doctor and a firefighter in the left section of the second output.

While Flux offers more robust results compared to SDXL, it still struggles with identity blending and incorrect spatial associations between individual text prompts and pose inputs. In contrast, FineControlNet successfully aligns text prompts with specific image regions and preserves the distinct identities described in the prompts, as illustrated in Figure 4 and 5. As FineControlNet is a training free method for diffusion-based image generation models, we believe that applying our method to newer models will allow users to continually achieve better fidelity while maintaining distinct identities for complex scenes.

9. Limitations

Despite showing promising results, our method can sometimes suffer from several failure modes, which include: 1) instance-specific controls being affected by the setting description, 2) human faces synthesized with poor quality, 3) implausible environments for the specified poses, and 4) misaligned input poses and generated images. The results of FineControlNet showing these failures are presented in Figure 11.

We observe that instance controls may get altered by the text prompt for the setting, especially in environments with small diversity of instances in the training dataset of images used for Stable Diffusion [12]. In addition, similar to ControlNet [14], our method can synthesize human faces that look unrealistic. We also can see unrealistic pairings of instances and environments in some of the generated images by FineControlNet. Even when satisfying the instance and setting specifications separately, our method can generate physically implausible scenes, such as floating people, as it does not have an explicit mechanism that prevents from doing so. Finally, FineControlNet can generate images whose poses are misaligned or with bad anatomy, particularly when the input poses are challenging.

Regarding the face quality, we conjecture the limitation is related to discretization errors due to downsampled latent embeddings in UNet during reverse diffusion. It can be also observed in other methods as shown in the main text' s qualitative comparison. The face quality improves by in-

[1] https://huggingface.co/docs/diffusers/en/index
[2] https://github.com/damian0815/compel

creasing scale of the input skeleton as shown in Figure 12. Alternatively, the limitation could be resolved by replacing the text-to-image backbone with an improved version.

10. Societal Impact and Ethical Concern

The deployment of Text-to-Image (T2I) models such as Stable Diffusion raises several ethical concerns that must be carefully considered. Firstly, there is the issue of content generation and potential misuse. These models can produce highly realistic images based on textual descriptions, which can be exploited to create deepfakes or other forms of misleading or harmful content. This poses significant risks in terms of misinformation, privacy violations, and reputational damage. Additionally, the training datasets for these models often contain biased or inappropriate material, which can inadvertently lead to the generation of biased or offensive images, perpetuating stereotypes or cultural insensitivity. Another concern is the potential for these models to be used inappropriately in contexts requiring high levels of trust and authenticity, such as in news media or legal evidence. Furthermore, the environmental impact of training large-scale AI models cannot be overlooked, as they require substantial computational resources and energy. Addressing these ethical issues involves implementing robust content moderation mechanisms, ensuring diversity and fairness in training datasets, and promoting transparency and accountability in the use of such technologies. Researchers and developers must engage in continuous dialogue with ethicists, policymakers, and the public to navigate these challenges responsibly.

References

[1] Stability AI. Sdxl 1.0. https://stability.ai/news/stable-diffusion-sdxl-1-announcement, 2024. 3, 8

[2] John Canny. A computational approach to edge detection. *TPAMI*, 1986. 2, 3

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 5

[4] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Kpe: Keypoint pose encoding for transformer-based image generation. *arXiv preprint arXiv:2203.04907*, 2022. 5

[5] Hugging Face. Diffusers multicontrolnet. https://github.com/huggingface/diffusers/tree/multi_controlnet, 2023. 5, 6

[6] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In *AAAI*, 2022. 2, 3

[7] Xuan Ju, Ailing Zeng, Wang Jianan, Xu Qiang, and Zhang Lei. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), year=2023*. 5

[8] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 5

[9] Black Forest Labs. Flux.1. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024. 3, 8

[10] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 5

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 8

[13] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 2, 3

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 5, 6, 8, 10, 11

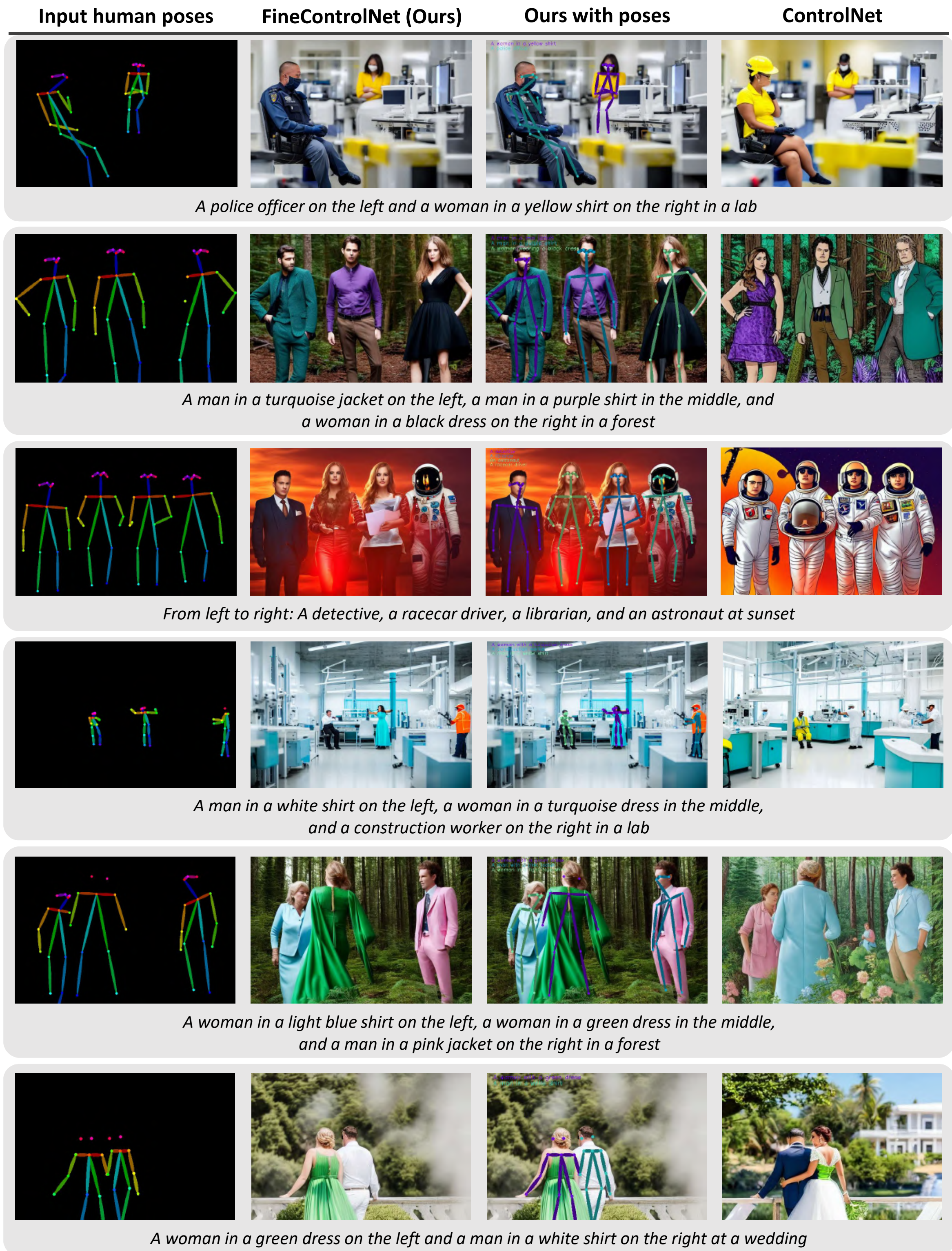


Figure 6. Additional supplementary results demonstrating our method's ability to finely control each instance in the image. We show the input poses (left) and prompt (bottom) along with the results from our method with and without overlaid poses (middle), and ControlNet's [14] output with the same text prompt (right) for comparison.

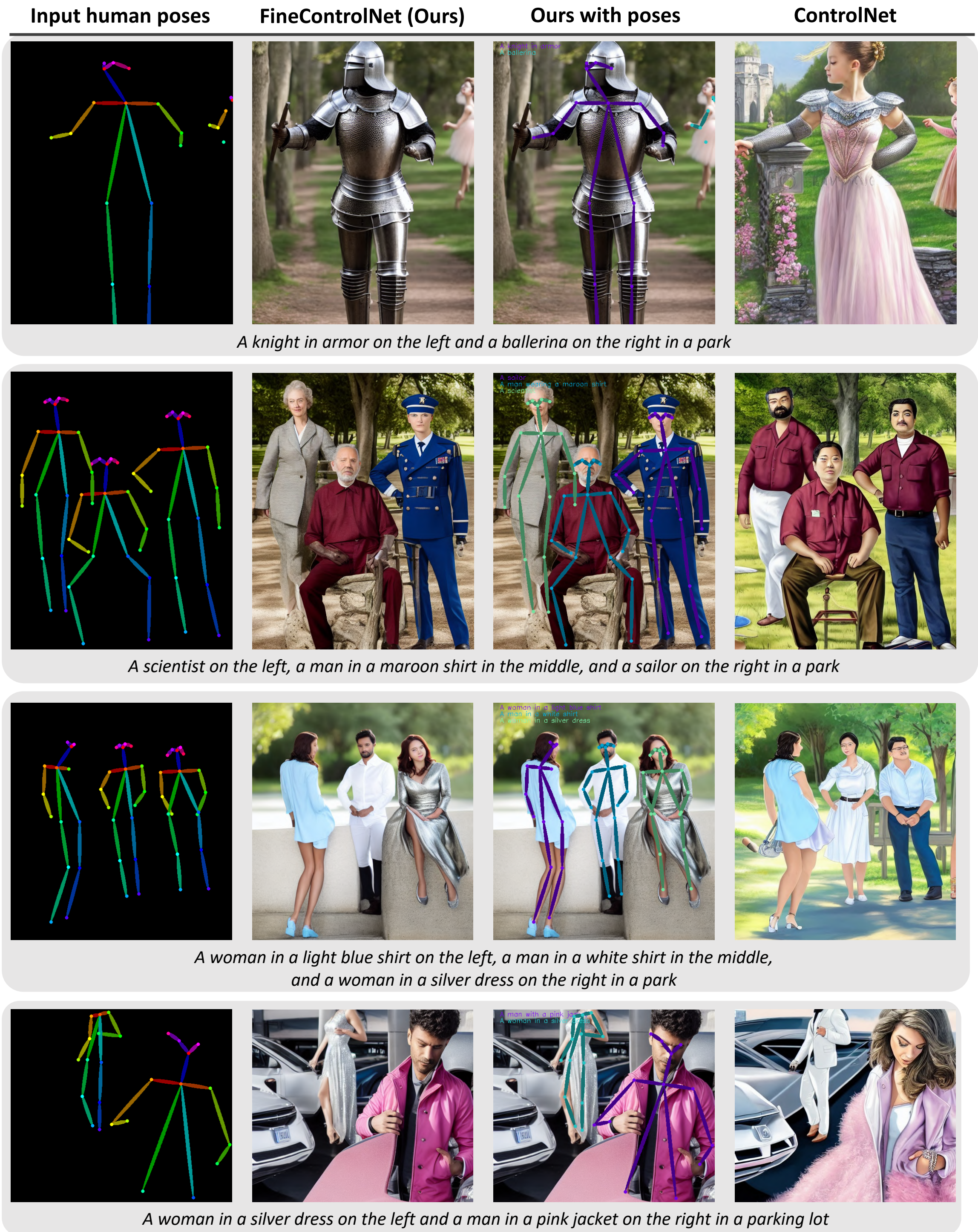


Figure 7. Additional supplementary results demonstrating our method's ability to finely control each instance in the image. We show the input poses (left) and prompt (bottom) along with the results from our method with and without overlaid poses (middle), and ControlNets's [14] output with the same text prompt (right) for comparison.

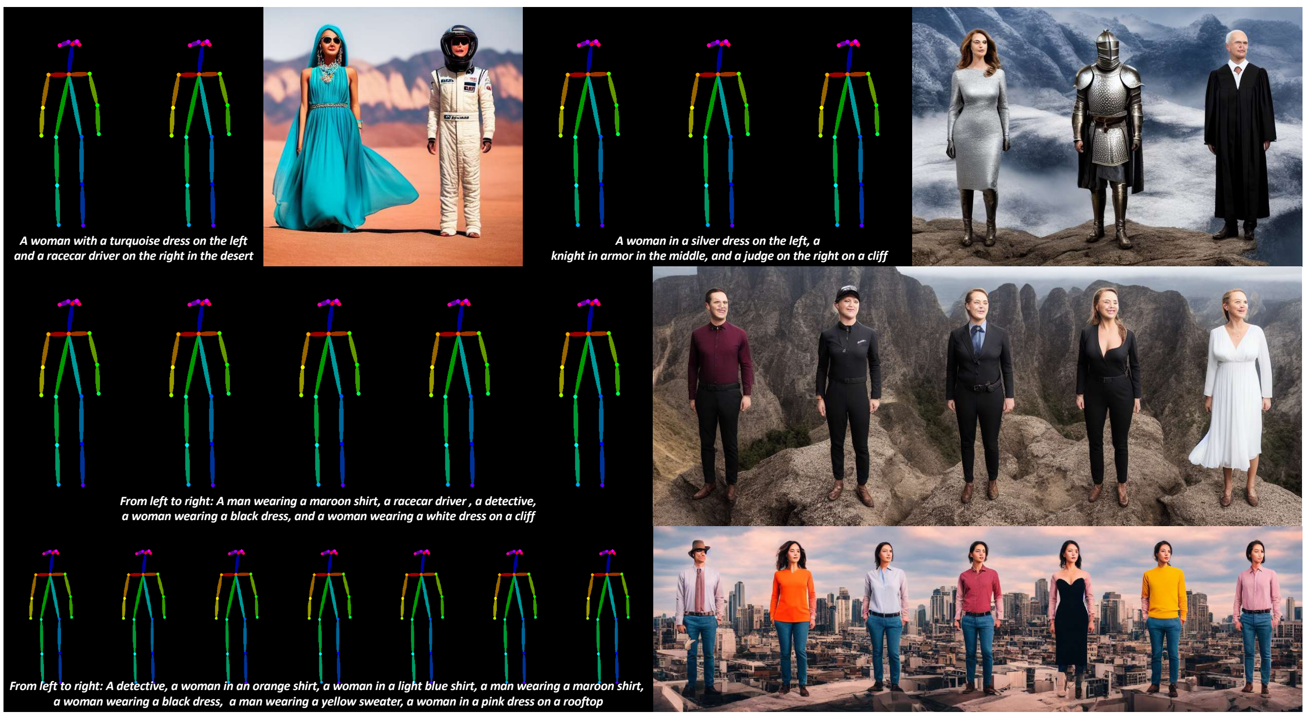
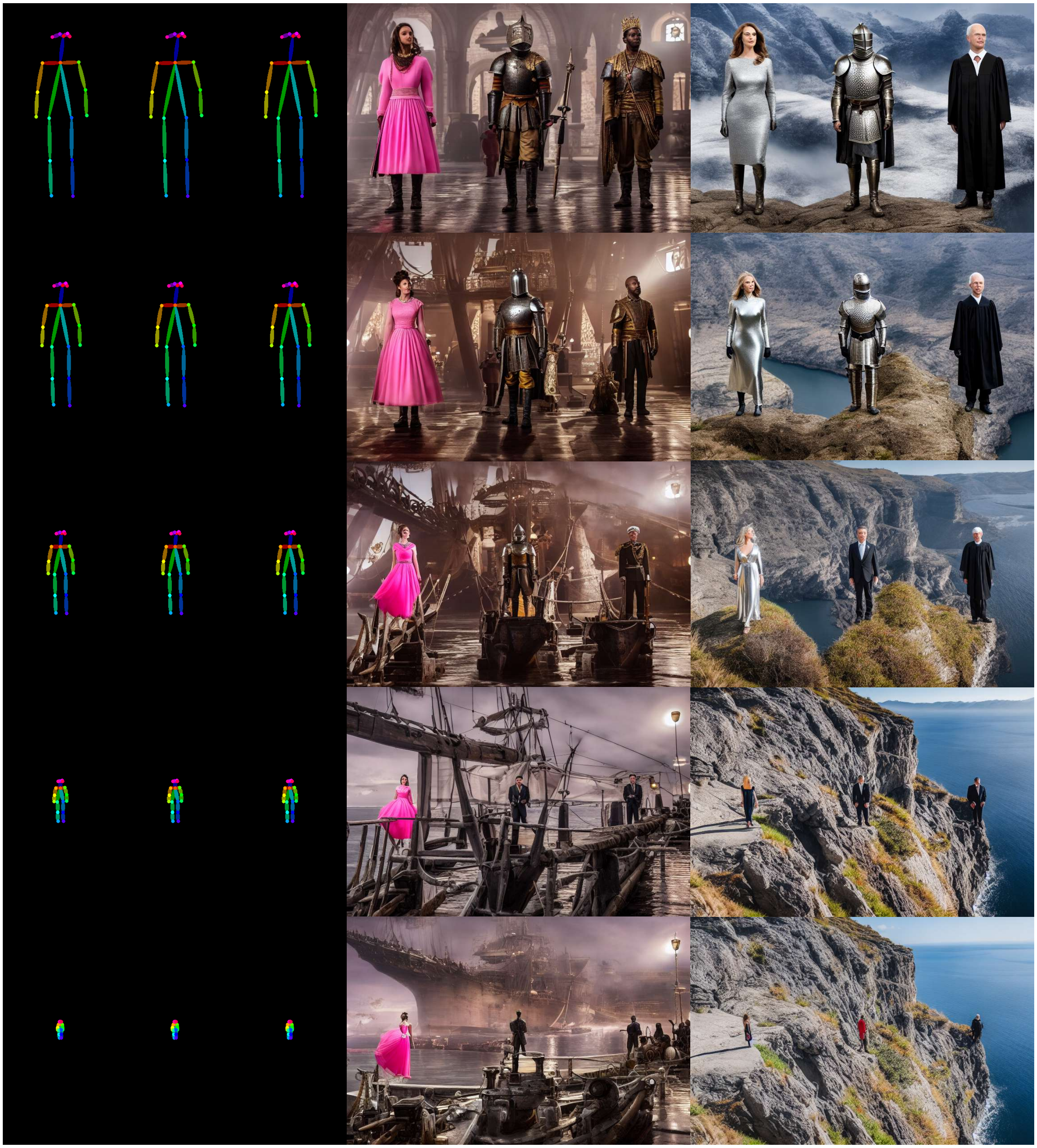


Figure 8. Qualitative results depending on the number of people, which is the number of 2D poses given. Every 2D human pose in the entire figure has the same resolution. The input skeleton map with 7 poses is resized to match the page.



From top to bottom, scales of 2D pose skeletons are 1.0, 0.75, 0.5, 0.25, and 0.1

A woman in a pink dress on the left, a knight in armor in the middle, and a king with a crown on a ship

A woman in a silver dress on the left, a knight in armor in the middle, and a judge on the righton a cliff

Figure 9. Qualitative results depending on the scale of a person, which represents the relative resolution of each pose in the input. We used the same seed for image generation for every scale variation.

From top to bottom, distances between 2D pose skeletons are 1.0, 0.75, 0.5, 0.25 in normalized scale

A racecar driver on the left, an astronaut in the middle, and a woman wearing a black dress on the right at a bar

A woman in a silver dress on the left, a knight in armor in the middle, and a judge on the righton a cliff

Figure 10. Qualitative results depending on the distance between people. Closer distance could cause *blending* between different instances' text embeddings and generate mixed appearance of instances. We used the same seed for image generation for every inter-personal distance variation.

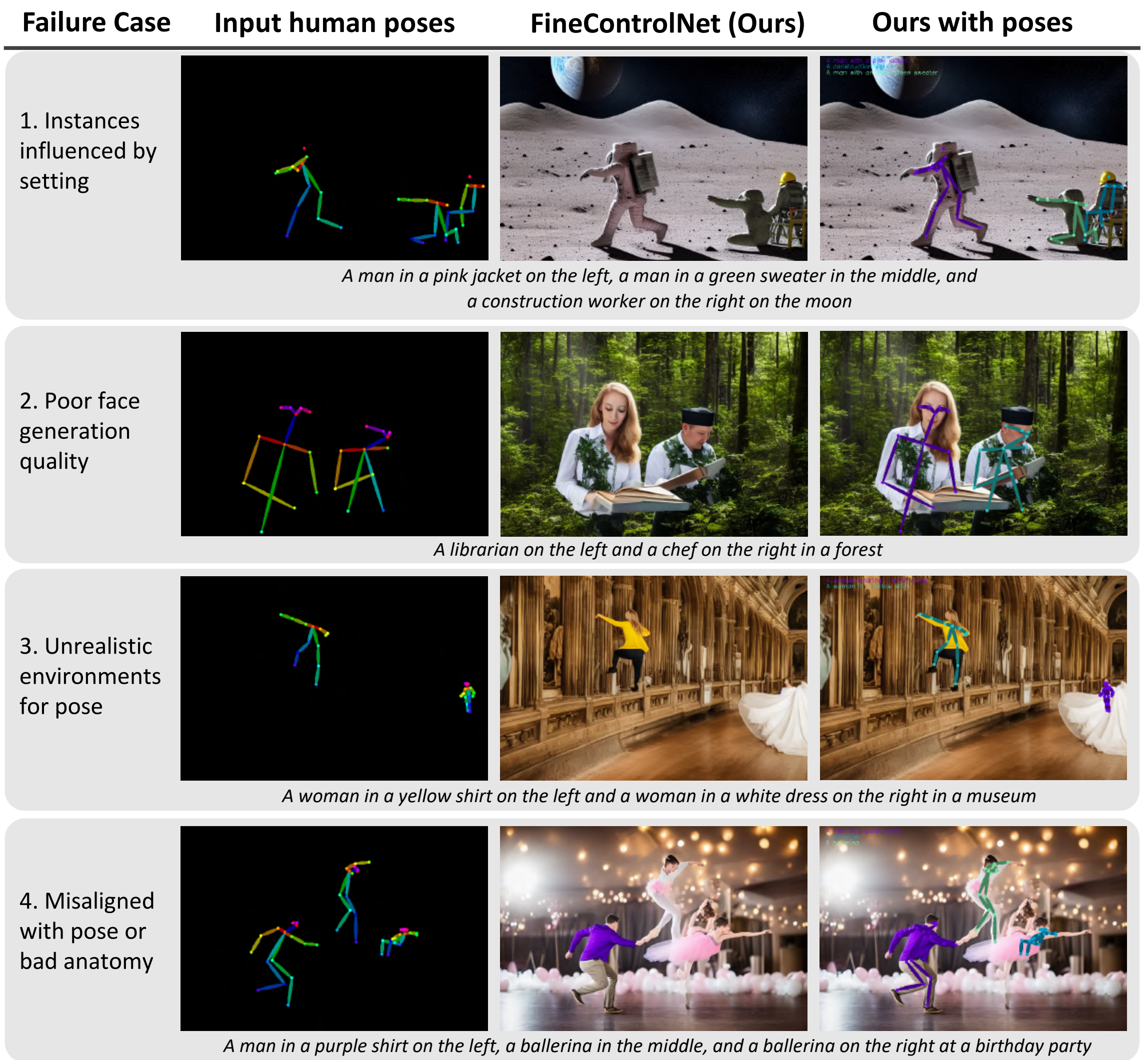


Figure 11. Failure cases. We demonstrate possible failure cases of FineControlNet that will be further studied in future work.

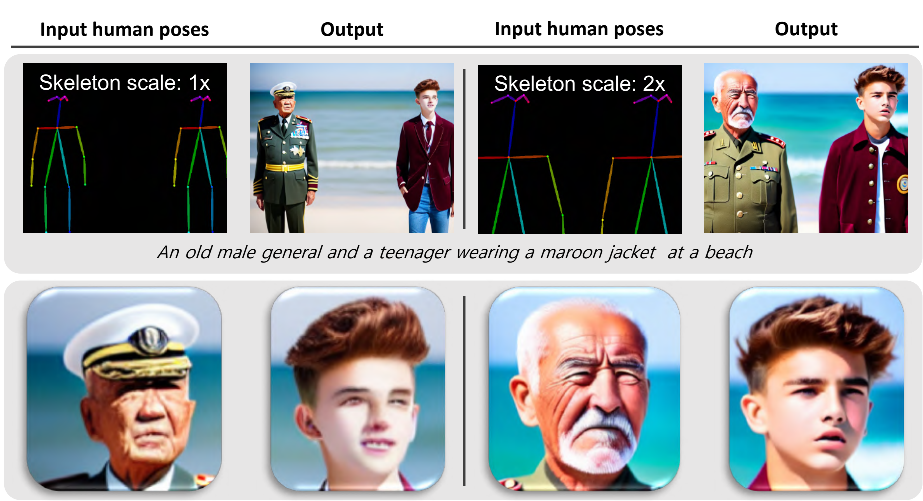


Figure 12. Face quality could be improved by increasing the scale of the input skeleton.