Supplementary of Rethink Sparse Signals for Pose-guided Text-to-image Generation

Wenjie Xuan[1], Jing Zhang[1]*, Juhua Liu[1]*, Bo Du[1], Dacheng Tao[2]
[1] School of Computer Science, Wuhan University [2] Nanyang Technological University
{dreamxwj,liujuhua,dubo}@whu.edu.cn, {zhangjing.cv,dacheng.tao}@gmail.com

Appendix

This appendix is organized as follows.

- Implementation details on the model structure. (§A)
- More details about datasets and keypoint definition. (§B)
- More discussions are provided, including the effects of loss coefficient η, an analysis of the decrement on the CLIP-Score, a validation of our heatmap constraint $\mathcal{L}_{ht}$, and effects of our method on data augmentation. (§C)
- Additional visualized examples with detailed illustrations are presented to supplement the main paper. (§D)
- A discussion about limitations and future works. (§E)

A. More Details of Model Structure

We provide the detailed structure of our sparse-pose embedding module $\mathcal{G}(\cdot)$ in Fig. S1. It comprises two linear layers and three basic blocks of stacked `Linear + GeLU + Dropout + Linear + Layer Norm` layers. The module accepts the random initialized vector $\boldsymbol{E}_0$ and outputs the learned keypoint embeddings $\boldsymbol{E}_{kpt}$ for constructing the spatial-pose representation, which is optimized directly for the denoising diffusion objective.

*Corresponding author

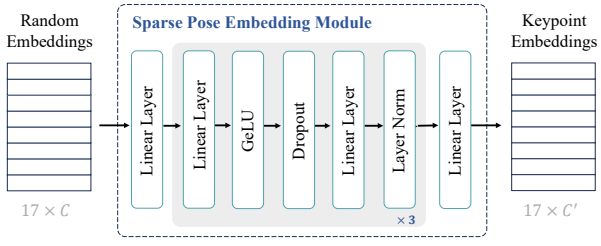


Figure S1. The detailed structure of our sparse pose embedding module $\mathcal{G}(\cdot)$.

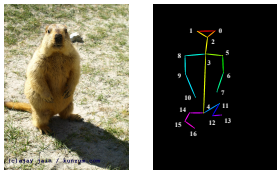
ID	Reference (main paper)	Brief Illustration
Fig. S1	§**3.1**, Line 191	Detailed architecture of the spatial embedding module $\mathcal{G}(\cdot)$.
Fig. S2	§**3.1**, Line 171 §**4.1**, Line 291,301	Keypoint descriptions of AP-10K and Human-Art dataset.
Fig. S3	§**4.2**, Line 328	Discussions on loss coefficient η.
Fig. S4	§**4.4**, Line 428	Analysis of the CLIP-Score.
Fig. S5	§**4.5**, Line 479	Discussion on heatmap loss $\mathcal{L}_{ht}$.
Fig. S6	–	Illustrations on the limitations.
Fig. S7	§**3.2**, Fig. 4	Full cross-attention maps of ControlNet.
Fig. S8	§**4.5**, Fig. 9	Full cross-attention maps of our SP-Ctrl.
Fig. S9	§**4.3**, Fig. 5	More examples on AP-10K.
Fig. S10	§**4.3**, Fig. 5	More examples on Human-Art.
Fig. S11	§**4.5**, Fig. 10	More examples of generation with different conditions.
Fig. S12	§**4.5**, Fig. 11	More examples to show the shape diversity of our method.
Fig. S13	§**4.5**, Fig. 12	More examples to show the cross-species generation.
Fig. S14	§**5.5**, Fig. 13	More examples to show the pose-editing results.
Tab. S2	§**4.1**, Line 295	Prompt templates for AP-10K.
Tab. S3	–	Validation for data augmentation.

Table S1. Quick overview of figures and tables in the Appendix.

B. More Details of Datasets

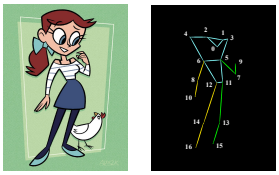
Definition of animal and human pose. Fig. S2 presents the definition of pose on the AP-10K [6] and Human-Art [3] datasets, including pre-defined keypoint descriptions and the topological skeletons. For AP-10K, we adopt the 17-keypoint definition of pose for mammals, which is provided by the dataset. For Human-Art, the dataset provides two kinds of pose definitions for the human, *i.e.*, 17-keypoints and 21-keypoints. To keep close to the definitions on ani-

(a) AP-10K



Keypoint	Description	Keypoint	Description
1	left eye	10	right elbow
2	right eye	11	right front paw
3	nose	12	left hip
4	neck	13	left knee
5	root of tail	14	left back paw
6	left shoulder	15	right hip
7	left elbow	16	right knee
8	left front paw	17	right back paw
9	right shoulder		

(b) Human-Art



Keypoint	Description	Keypoint	Description
1	nose	10	left wrist
2	left eye	11	right wrist
3	right eye	12	left hip
4	left ear	13	right hip
5	right ear	14	left knee
6	left shoulder	15	right knee
7	right shoulder	16	left ankle
8	left elbow	17	right ankle
9	right elbow		

Figure S2. The definition (or description) of each keypoint on the AP-10K and Human-Art dataset.

Table S2. Prompt templates for the AP-10K dataset, where `<CLS>` and `<BG>` denote specie names and background types.

ID	Prompt Templates
1	`A good photo of <CLS>.`
2	`A photo of <CLS> in the <BG>.`
3	`There is <CLS> on the <BG>`
4	`There are some <CLS> lying in the <BG>.`
5	`Some <CLS> are in the <BG>.`
6	`A close photo of <CLS>.`
7	`In the <BG>, there are several <CLS>.`
8	`This is a clear photo of <CLS> in the <BG>.`
9	`Several <CLS> are on the <BG>.`
10	`A <CLS> stands on the <BG>.`

mal poses, we employ the 17-keypoint definition of human pose as listed in Fig. S2. To visualize the animal and human pose in OpenPose style, we utilize the code provided by the popular `mmpose` [1] repository.

Prompt Templates for AP-10K. Since the AP-10K dataset does not provide the image captions, we utilize a series of prompt templates following common practice [2]. We designed 10 templates as presented in Tab. S2, where the species name and background category are utilized to construct the textual prompt for each image. The AP-10K dataset predefines 54 species of 23 animal families and 8 background types [6], including *grass or savanna, forest or shrub, mud or rock, snowfield, zoo or human habitation, swamp or rivderside, desert or gobi, and mugshot*. When training, we randomly select one as the image caption.

C. More Discussions

Discussions of the loss coefficient η. We search for the optimal setting of the loss weighting η for our proposed heatmap loss $\mathcal{L}_{ht}$ as reported in Fig. S3. While too large

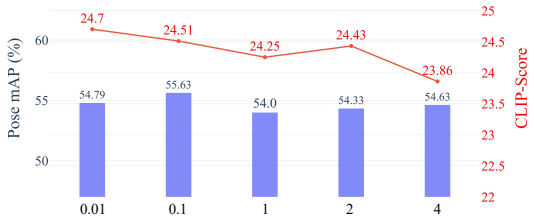


Figure S3. Discussions of the loss coefficient η.

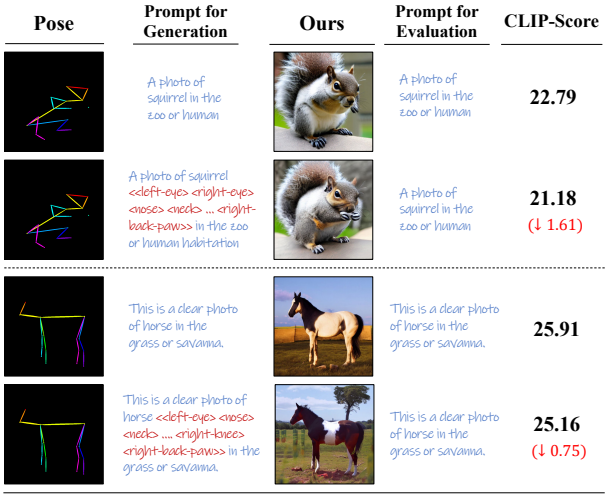


Figure S4. An illustration on the decrease of CLIP-Score due to inference-evaluation discrepancy.

η decreases the CLIP-Score and pose mAP, we set $\eta = 0.1$ in our experiments, which achieves the optimal pose mAP of 55.63% and competitive CLIP-Score of 23.86.

Explanations on the computation of CLIP-Score. Since our method binds text prompts with newly introduced keypoint tokens in the main paper §**3.2** (denoted as `<kpt>`) for training and generation, this arises difficulty in properly evaluating CLIP-Score of `<kpt>`. The reasons are: 1) `<kpt>`s are new to the pretrained CLIP model, which are not contained in the CLIP model's vocabulary, and 2) fine-tuning the CLIP model with `<kpt>` would also bring biases. Thus, we take their initialized tokens as alternatives to compute the CLIP-Score for estimation, approximating the true CLIP-Score as possible. The recomputed CLIP-Score 24.51 is comparable to 24.77 of the baseline ControlNet in the main paper Fig. 7, and the difference is almost negligible. Besides, we suppose that it is possible to obtain a trade-off between the pose mAP and CLIP-Score by dropping `<kpt>` during parts of the sampling steps. We also tried to simply remove all `<kpt>` for estimation, but found decreased CLIP-Scores as illustrated in Fig. S4. It is considered an error caused by inference-evaluation discrepancy.

Effects of the heatmap constraint $\mathcal{L}_{ht}$. To validate the

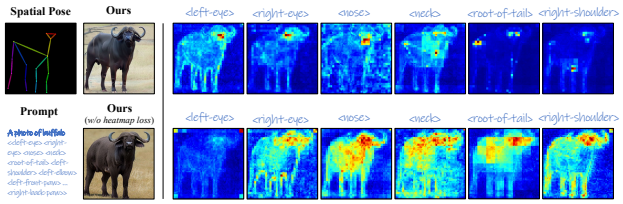


Figure S5. Comparisons of visualized cross-attention maps with and without the heatmap constraint $\mathcal{L}_{ht}$.

Data Gen	mAP↑	mAR↑
ControlNet	50.39	55.17
Ours	**57.10**	**61.00**
Real Image	74.70	77.68
+ ControlNet	75.05 (+0.35)	78.17 (+0.49)
+ Ours	**75.44** (+0.74)	**78.49** (+0.81)

Table S3. Validation on the effectiveness of our method as a data generator on AP-10K for data augmentation.

function of $\mathcal{L}_{ht}$, we visualize the cross-attention map with and without the proposed heatmap constraint. As shown in Fig. S5, without explicit constraints on the heatmaps, the newly introduced tokens are spatially overlapped and lack specific meanings, failing to attend to the correct positions of keypoints. In contrast, the heatmap constraint $\mathcal{L}_{ht}$ effectively distinguishes attentions of different keypoint tokens and benefits pose accuracy in the main paper, Fig.7. Moreover, even though exploring other solutions for constraints is feasible, the heatmap constraint is straightforward and reasonable for pose control, and it shows no obvious defects in image generation and pose control. Thus, we adopt heatmaps to encourage sparse and strong spatial responses to each keypoint for enhanced pose alignments. Note that the heatmaps are Gaussian-blurred [4].

Validation of our method as a data generator for augmentation. To prove the effectiveness of our method, we evaluate it as a data generator. Specifically, we generate images from pose annotations of AP-10K and train *HRNet* on the synthesized images for pose estimation. As reported in Tab. S3, when employing the synthesized images only or combining them with real images for training, our method outperforms ControlNet under both settings and improves *HRNet* with no special designs except data augmentations, which obtains similar gains as observed in [5]. This fact indicates better pose alignments than ControlNet and unravels the potential in data augmentation.

D. More Visualized Results

We provide additional qualitative results to supplement the main paper. Details are as follows.

Full visualized results of cross-attention maps. We show all the cross-attention maps of the vanilla ControlNet [7] and our *SP-Ctrl* method in Fig. S7 and Fig. S8, where the time step here denotes the steps of adding noises. These attention maps are averaged across different transformer blocks at each time step. As shown in the figure, compared with the baseline ControlNet, the keypoint tokens attend to the positions of each keypoint more accurately. Though we only compute the $\mathcal{L}_{ht}$ among the 3^{rd} transformer blocks during the 250∼500 time steps following the best practice, we notice that the cross-attention maps after the 250∼500 time steps also attend to the keypoint positions. This fact indicates that constraining the cross-attention maps during the 250∼500 time steps implicitly regularizes the attention maps at other time steps, all contributing to the learning of new keypoint tokens.

More visualized comparisons between other popular methods and ours. Here we present more examples of pose-guided image synthesis on the AP-10K and Human-Art dataset in Fig. S9 and Fig. S10 to show the effectiveness of our SP-Ctrl. As shown in the figure, while ControlNet and other methods might fail to interpret certain keypoints, such as the limbs, our method shows advantages in aligning with the detailed poses on both animal- and human-centric generation tasks. These results demonstrate the effectiveness of our method in pose controllable generation with sparse signals.

More visualized comparisons on different pose guidance. Fig. S11 presents more visualized examples of ControlNet with different conditions and ours. The masks fail to control the keypoint positions. While the depth map achieves precise control over pose, it also constrains the shape of generated animals. The OpenPose signal provides pose guidance for ControlNet but may fail when meeting complex poses or overlapped local structures. In contrast, our method achieves better control over sparse pose signals.

More examples to show the advantages of our method. Benefiting from the precise pose control under the sparse pose guidance, our SP-Ctrl shows several appealing properties for applications. Compared to dense signals like depth, our method exhibits more diverse results in object shapes, as shown in Fig. S12. Moreover, due to the category-agnostic characteristics of sparse pose, our method enables cross-species generation. As shown in Fig. S13, despite the discrepancies in the action and skeleton proportions among different animal species, our method can produce promising results of different animals sharing the same pose. Additionally, since the sparse pose signals do not necessarily rely on the pretrained pose estimators, it enjoys great flexibility in pose editing and creation. Fig. S14 showcases several examples. Such results show the great potential of sparse pose signals in spatially controllable generation.

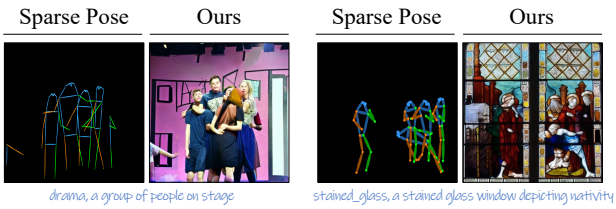


Figure S6. Examples of multiple-instance generation with pose signals, which is a more challenging case with overlaps, occlusion, and interactions of subjects.

E. Limitations and Future Work

By far, we have demonstrated the effectiveness of our SP-Ctrl in pose-guided text-to-image generation using sparse signals, which achieves performance nearly comparable to dense signal-based methods in terms of pose alignment. However, a significant gap in pose accuracy ($> 25\%$) remains between synthesized images and real ones, particularly for rare or complex poses. One possible solution is to leverage synthesized images to augment the pose diversity, particularly for complex ones, to enhance the perception and pose-alignment of generative models. Another crucial challenge is pose-controllable generation involving multiple instances. Although our method has shown promising results in Fig. S9 and Fig. S10, further researches are required to address the multiple-instance generation with pose signals, which is a more challenging case with overlaps, occlusions, and interactions between subjects as presented in Fig. S6. We leave this for future work.

References

[1] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[3] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 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*, 2023. 1

[4] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[5] Lihe Yang and et al. Synthetic images with dense annotations make stronger segmentation models. In *NeurIPS*, 2023. 3

[6] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 2

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

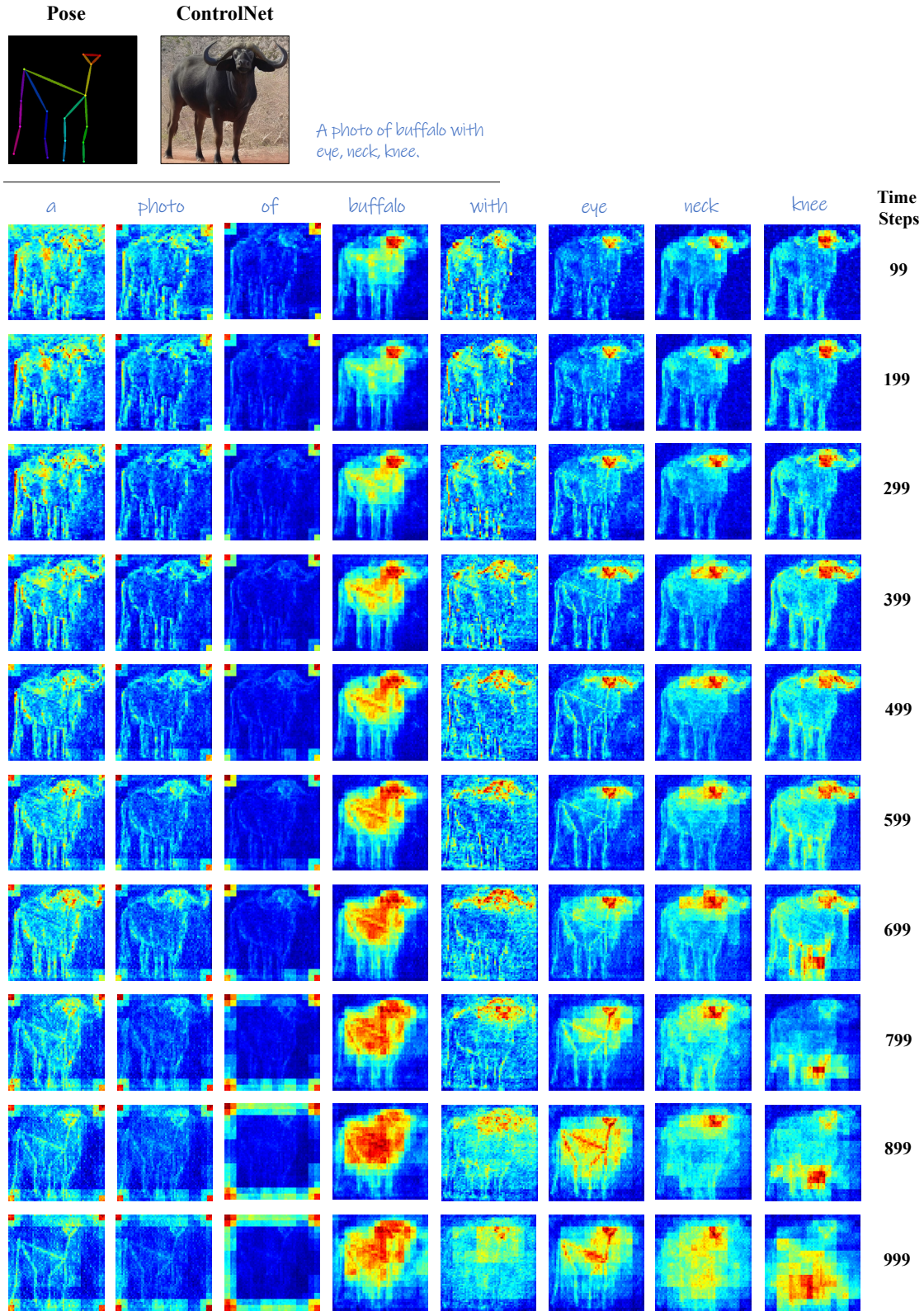


Figure S7. Visualized cross-attention maps of ControlNet at different time steps, which are averaged from all cross-attention layers. The time step here denotes the steps of adding noises.

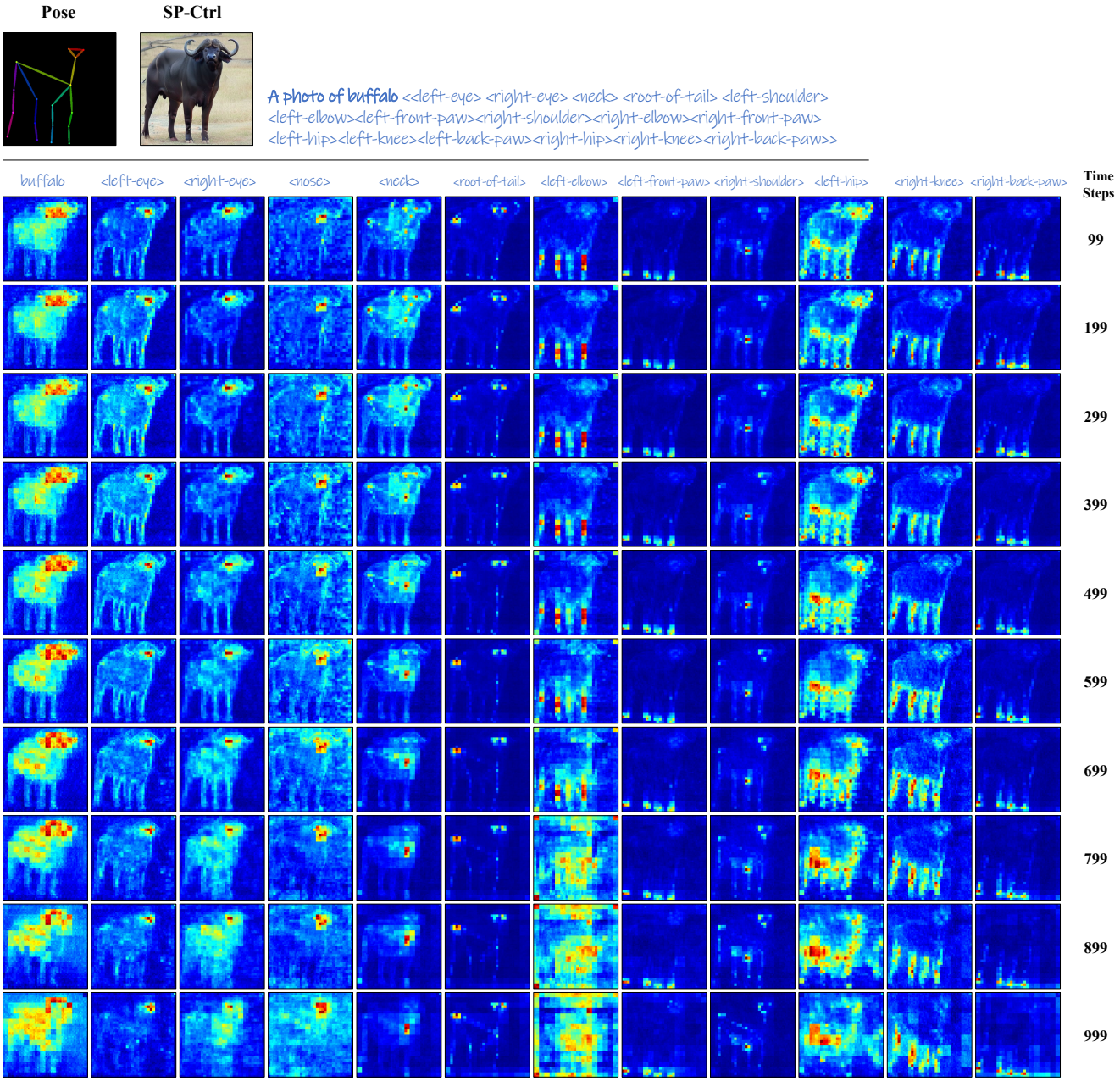


Figure S8. Visualized cross-attention maps of our *SP-Ctrl* at different time steps, which are averaged from all cross-attention layers. For page width limitations, we present the cross-attention maps of distinct keypoint for visualization. The time step here denotes the steps of adding noises.

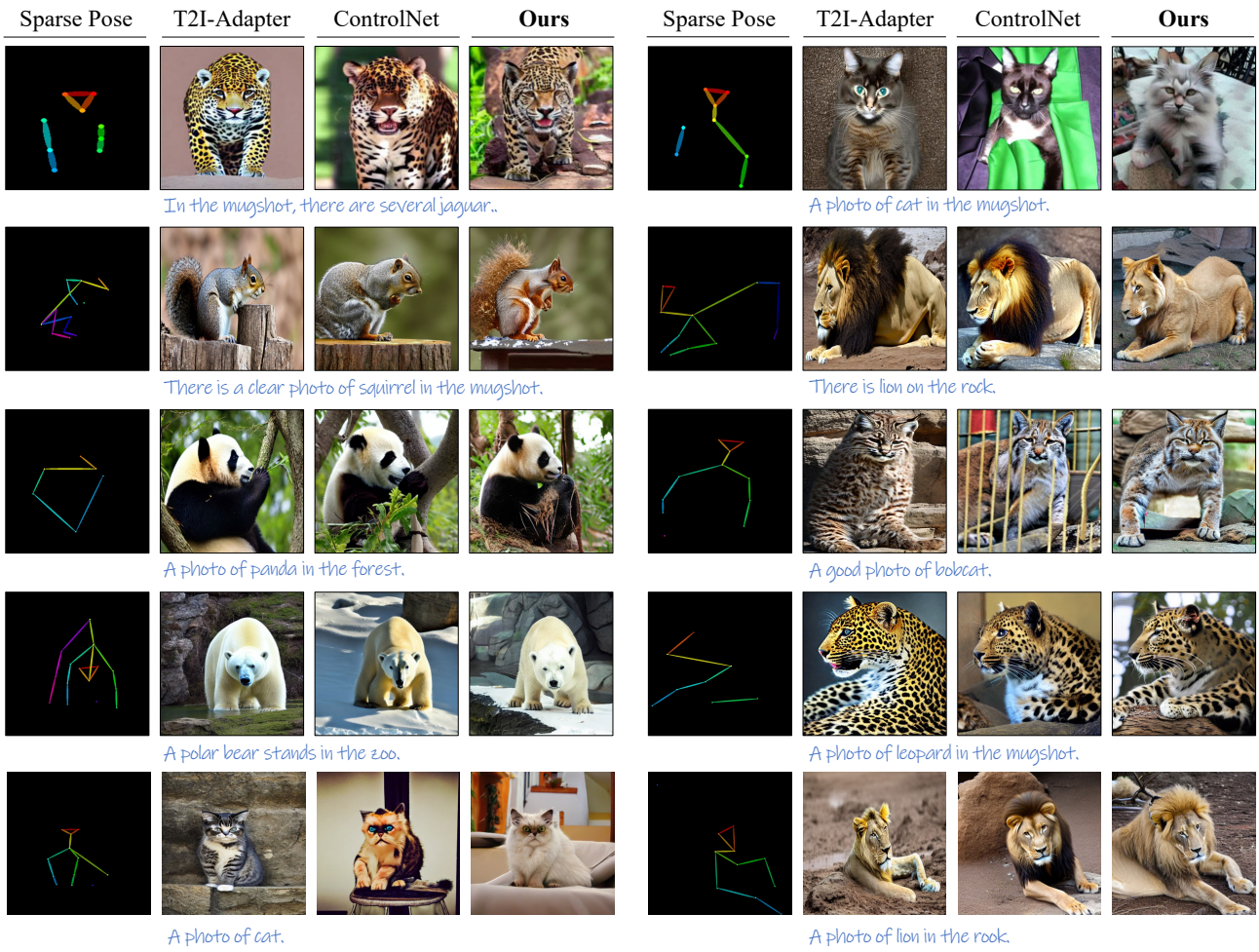


Figure S9. More visualized examples of our method and other popular methods on the AP-10K dataset.

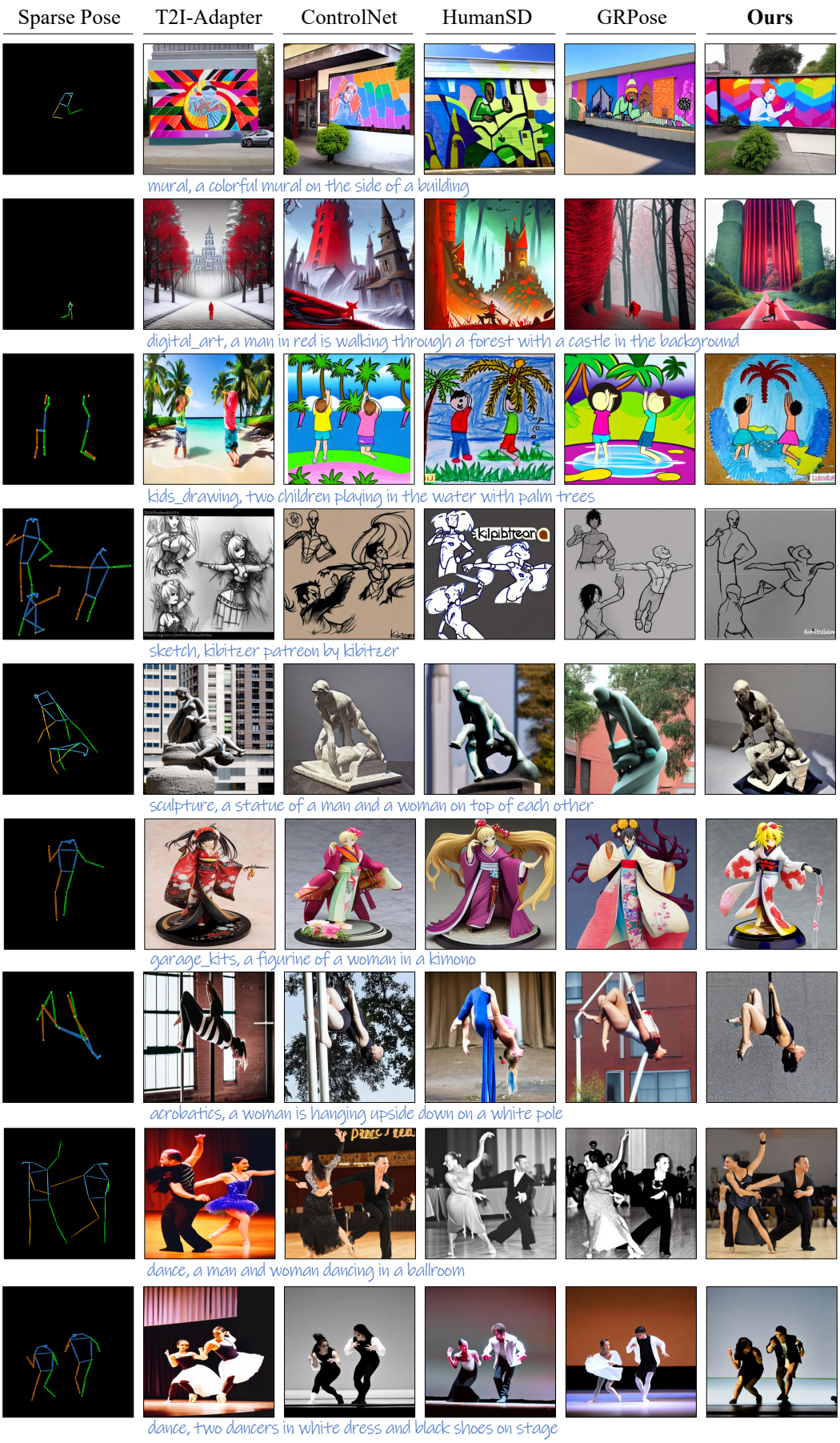


Figure S10. More visualized examples of our method and other popular methods on the Human-Art dataset.

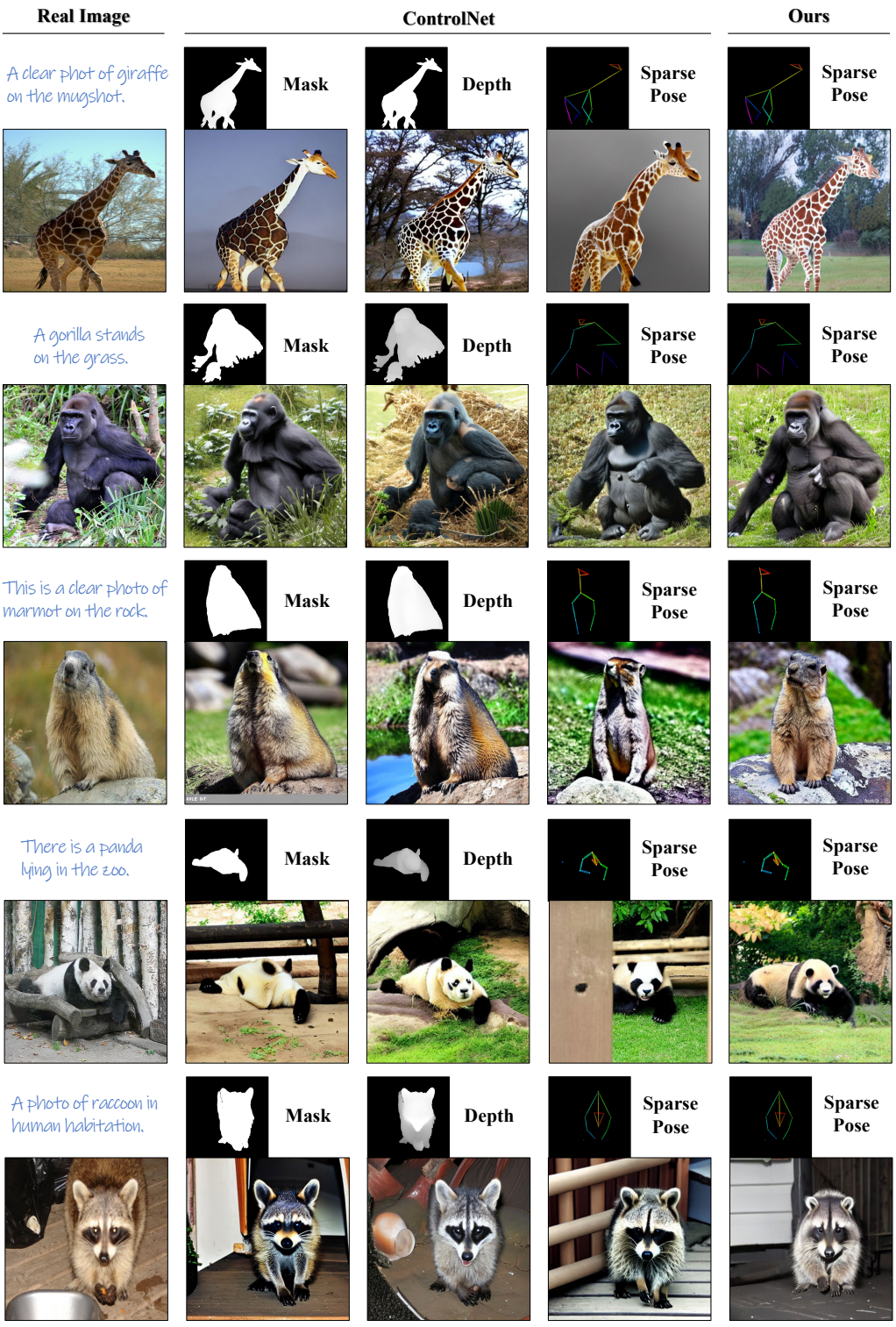


Figure S11. More visualized examples generated with different conditional guidance of ControlNet and our method.

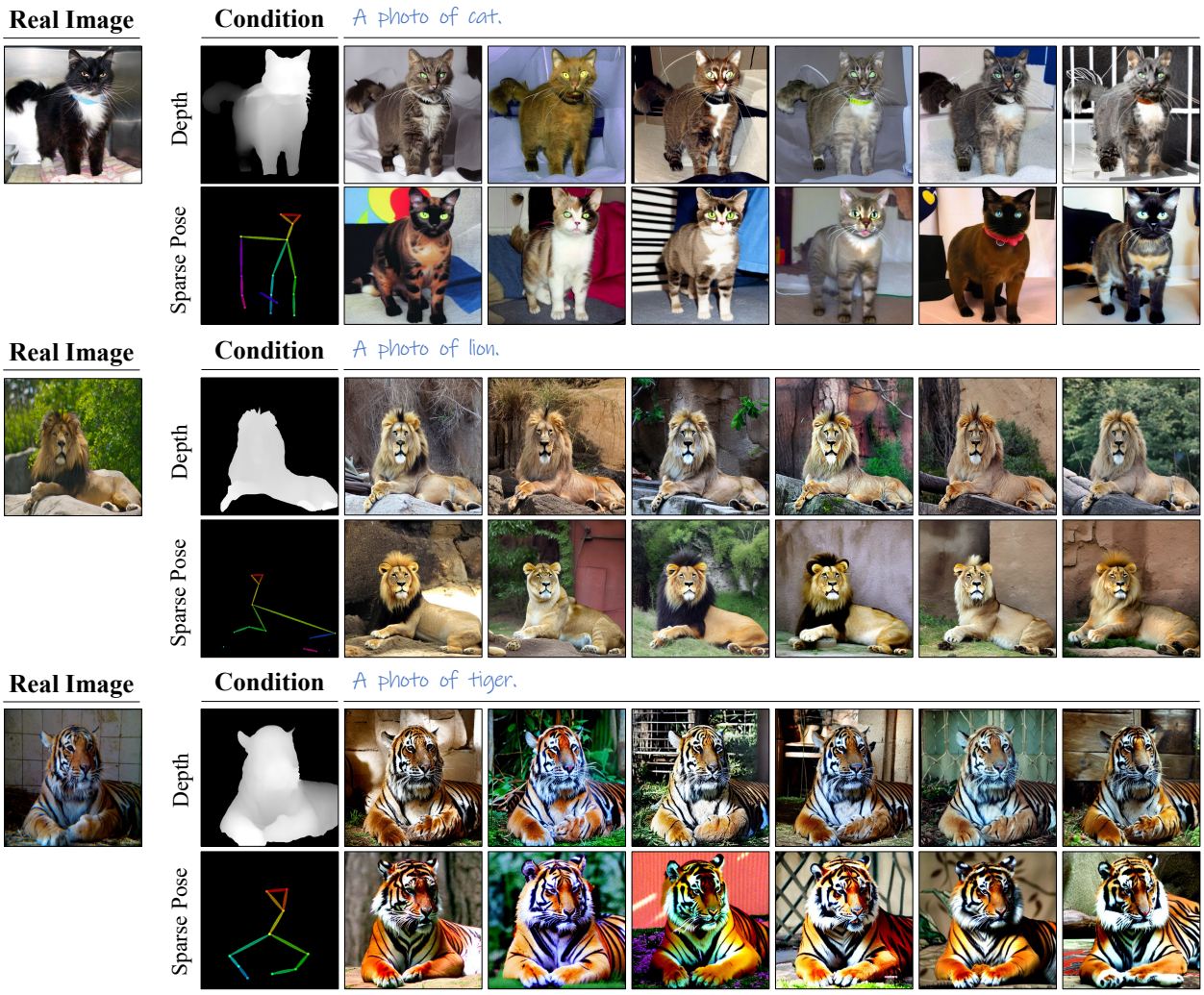


Figure S12. More visualized examples to showcase the shape diversity of synthesized images. When achieving precise pose control comparable to depth maps, our method can generate more diverse results with sparse signals, especially in shapes and contours.

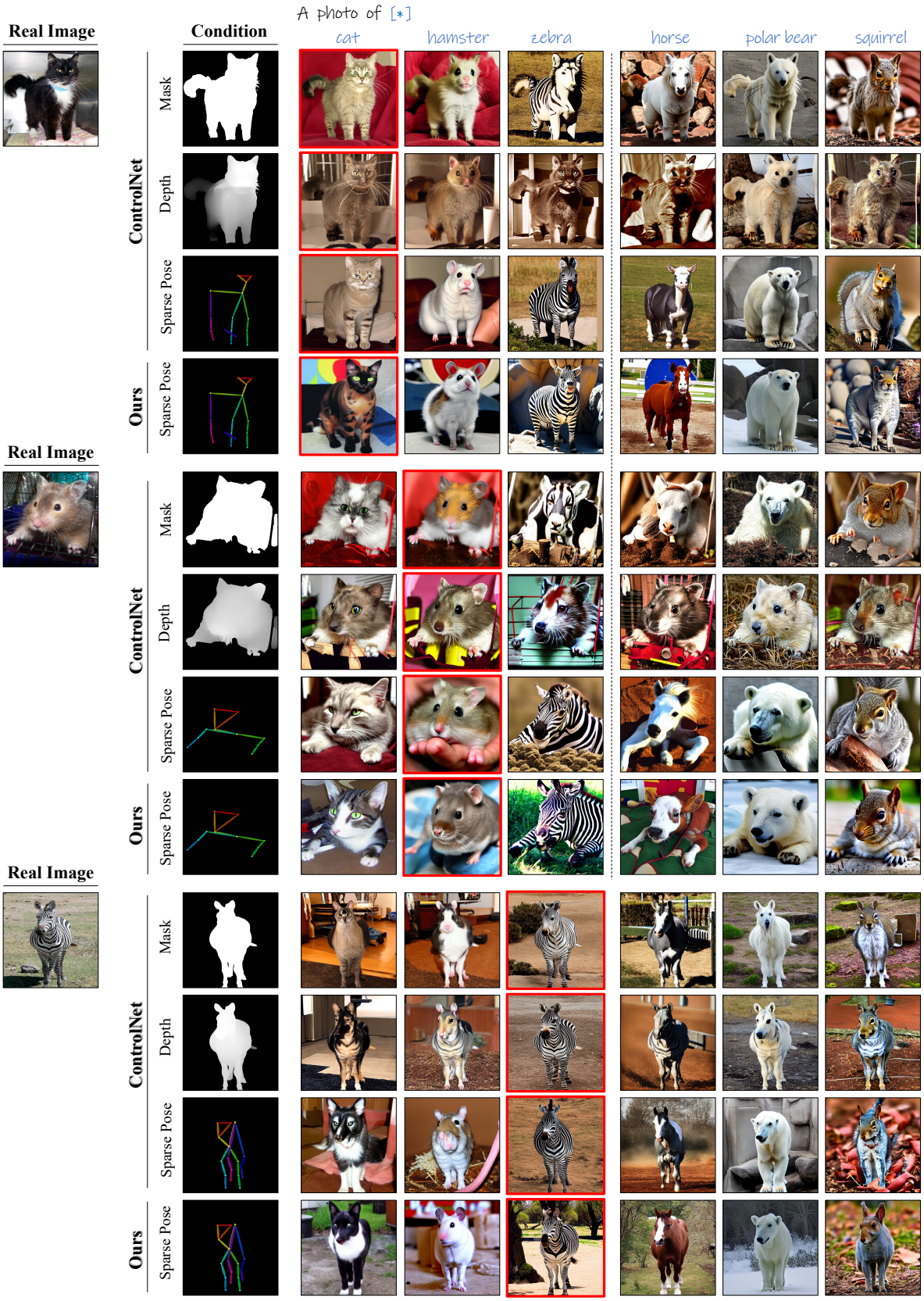


Figure S13. More visualized examples to showcase the cross-species generation ability. Our method can not only generate other species of animals but also keep high pose accuracy and image fidelity.

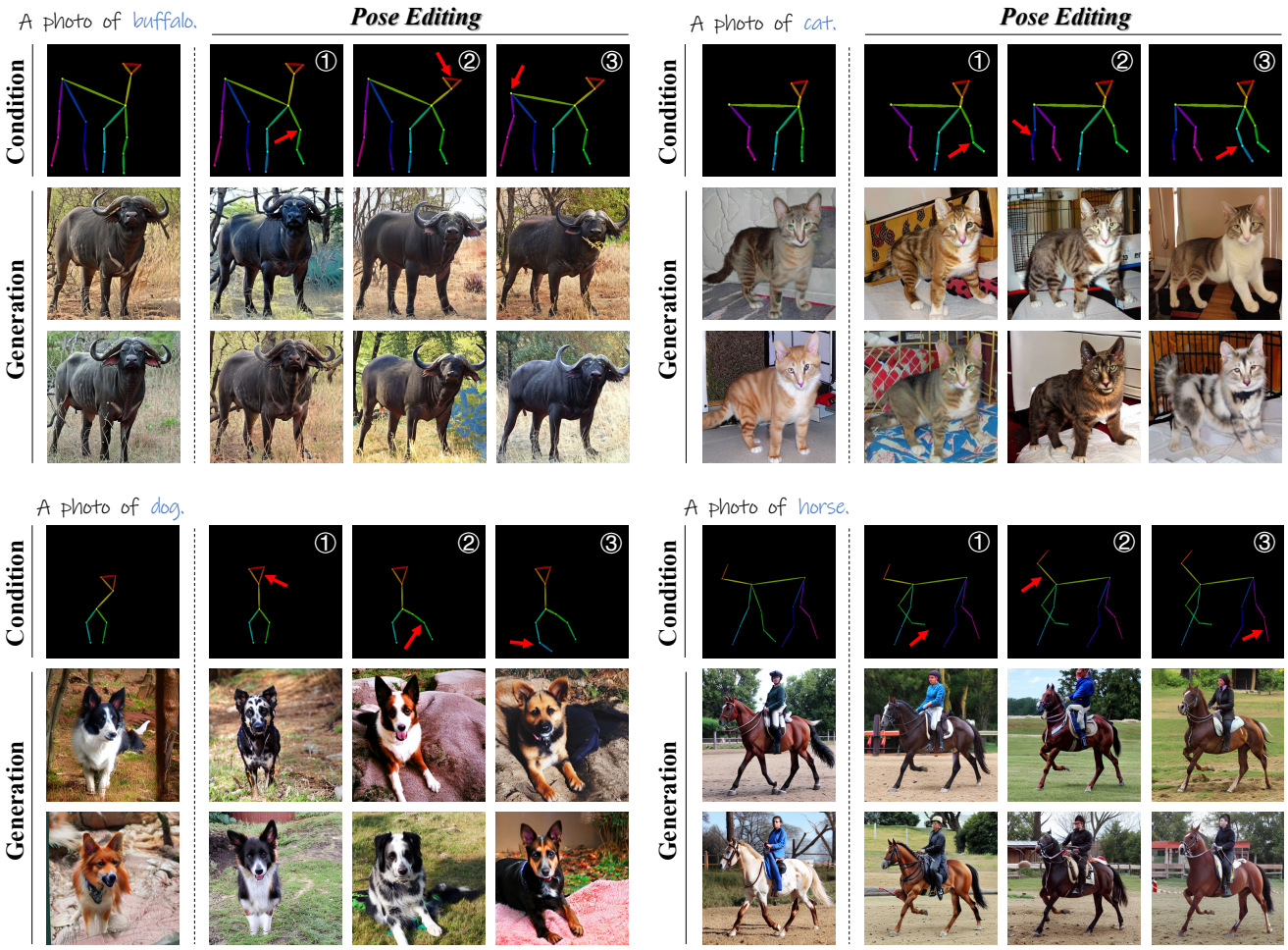


Figure S14. More visualized examples to showcase the generated image with edited pose signals, where sparse poses can be easily edited by changing the positions of keypoints.