# Skip-and-Play: Depth-Driven Pose-preserved Image Generation for Any Objects
*– Supplemental Document –*

Kyungmin Jo
KAIST
Daejeon, Korea
bttkm@kaist.ac.kr

Jaegul Choo
KAIST
Daejeon, Korea
jchoo@kaist.ac.kr

## A. Overview

In the supplementary document, we describe the experimental setups, including the datasets (Sec. B), evaluation metrics (Sec. C), and implementation details (Sec. D). Additionally, we provide a further analysis of ControlNet, examining its global weights (Sec. E) and evaluating its performance in terms of pose accuracy across different conditions and diffusion models (Sec. F), demonstrating that the proposed approach is not limited to specific conditions or diffusion models. Finally, we establish the superiority of our method by presenting additional results across various conditions and prompts in Sec. G.

## B. Datasets

In this section, we describe the datasets used in the experiments. We utilize images from the internet and four in-the-wild datasets to validate our method: FFHQ [11], AFHQ [5], CompCars [12], and LSUN Church [19]. FFHQ is employed for analyzing ControlNet [20] and for comparing the performance of SnP with pose-guided image generation models, while the remaining datasets are used to present qualitative results. Initially, we partition the FFHQ dataset into three distinct subsets: FFHQ-v, FFHQ-C, and FFHQ-B. FFHQ-v serves as a validation set specifically for analyzing the behavior of ControlNet, conducting ablation studies, and comparing pose accuracy between ControlNet conditioned on depth and keypoints, as discussed in Sec. 4, Sec. 5.4, and Sec. 5.5 of the main paper. FFHQ-C and FFHQ-B, on the other hand, are used as test sets for comparing the performance of SnP with baseline models in Sec. 5.1 of the main paper. To elaborate, we sample images for image prompts from FFHQ-C and use FFHQ-B as a real dataset for measuring the FID score, as detailed in Sec. C. The pose distributions across the three datasets are as follows: FFHQ-C replicates the centralized pose distribution of the original FFHQ dataset, while FFHQ-B adopts a uniform pose distribution. The poses in FFHQ-B and FFHQ-C cover rotation angles from $-50°$ to $50°$ and elevation angles from $-20°$ to $30°$. In contrast, FFHQ-v maintains a uniform rotation angle distribution but sustains a centralized elevation pose distribution, offering a broad range of angles. Specifically, FFHQ-v includes rotation angles from $-90°$ to $90°$ and elevation angles from $-55°$ to $45°$. Each dataset contains distinct images, ensuring no overlap. FFHQ-B and FFHQ-C each consist of 5,065 images, while FFHQ-v contains 100 images. For the other datasets, we use the provided test sets.

In addition, we construct two human pose datasets, PoseH-v and PoseH-B, for sampling pose images for analysis and evaluation, respectively. Recognizing that the FFHQ dataset has a limited range of poses and primarily consists of images captured at near-frontal angles, we create an additional pose dataset with a uniform and wide-ranging pose distribution. This enables a more accurate assessment of the effects of models on the pose of generated images across various angles. The poses in both the PoseH-v and PoseH-B datasets span rotation angles from $-90°$ to $90°$ and elevation angles from $-20°$ to $30°$, mirroring the pose distribution of FFHQ. Specifically, we follow the rotation range of FFHQ and the elevation range of FFHQ-B. This decision is based on the observation that most head pose estimation models exhibit lower accuracy in estimating elevation compared to rotation [9], especially as the pose deviates further from the frontal view [8]. To improve the accuracy of pose estimation, we exclude elevation angles within certain ranges that are infrequent in the FFHQ dataset. PoseH-v and PoseH-B contain 100 and 5,065 images, respectively. Each pose dataset consists of ground truth poses, two types of depth conditions, synthetic images, and two types of keypoints. The ground truth poses are uniformly sampled from the distribution. Synthetic images are generated using SD [15] to obtain conditions. One type of depth condition is rendered from a single 3D mesh of a human obtained from Objaverse [7], utilizing poses through Blender [6] for ControlNet, SmartControl [13], and SnP (SD). The other depth condition is estimated from synthetic images for ControlNet and SnP based on SDXL using ZoeDepth [1]. Additionally, two types of keypoints (KP) are obtained from OpenPose [3] and DIFT [17] for ControlNet-KP and DragDiffusion [16], respectively.

## C. Evaluation Metrics

In this section, we explain the metrics used for the quantitative performance comparison. We evaluate our method and baseline models based on pose accuracy, prompt alignment, and the photorealism of the generated images. To assess pose accuracy, we compare the estimated pose of generated images using an off-the-shelf pose estimation model [9] with the ground truth pose used during rendering. The pose discrepancy is quantified using the root mean square error (RMSE), measured in degrees. For prompt alignment, we compute the CLIP cosine similarity [14] between the generated images and the image prompts sampled from FFHQ-C. To evaluate photorealism, as detailed in [4], we compute the Frechet Inception Distance (FID) [10] between a combined set of 10,130 images from FFHQ-B, which includes images with x-flip applied, and the generated images. Although FFHQ-C is used for sampling image prompts, the FID score is computed using FFHQ-B for two reasons. First, FFHQ-B contains images not used as prompts, allowing us to assess photorealism without the influence of the image prompts. Second, since FID evaluates the discrepancy between the distribution of real images and generated images, it incorporates the pose distribution of the real dataset into the evaluation. FFHQ-B, with its pose-balanced distribution, is preferred over FFHQ-C, which is dominated by frontal-facing poses.

## D. Implementation Details

We employ 30 inference steps with a guidance scale $s$ set to 7.5 for SD and 5.0 for SDXL. Additionally, we configure $\lambda_t$ as 0.3 for all experiments. Also, we use a global weight of 0.6 for ControlNet and SnP in SDXL. This is based on the analysis in Sec. E, which shows that higher global weights in SDXL significantly reduce the prompt alignment. In SnP, in addition to utilizing the global weight, we optionally incorporate pixel-wise weight maps obtained from the Weight Control Module (WCM) to adjust the weights of ControlNet features, as outlined in Sec. 4.2. We specifically adjust the values of these weight maps to range between 0.5 and 1.0 for SD, and between 0.6 and 1.0 for SDXL. These ranges enable the model to maintain the pose of depth condition while enhancing the fidelity of shape reflection from the prompt (Sec. E).

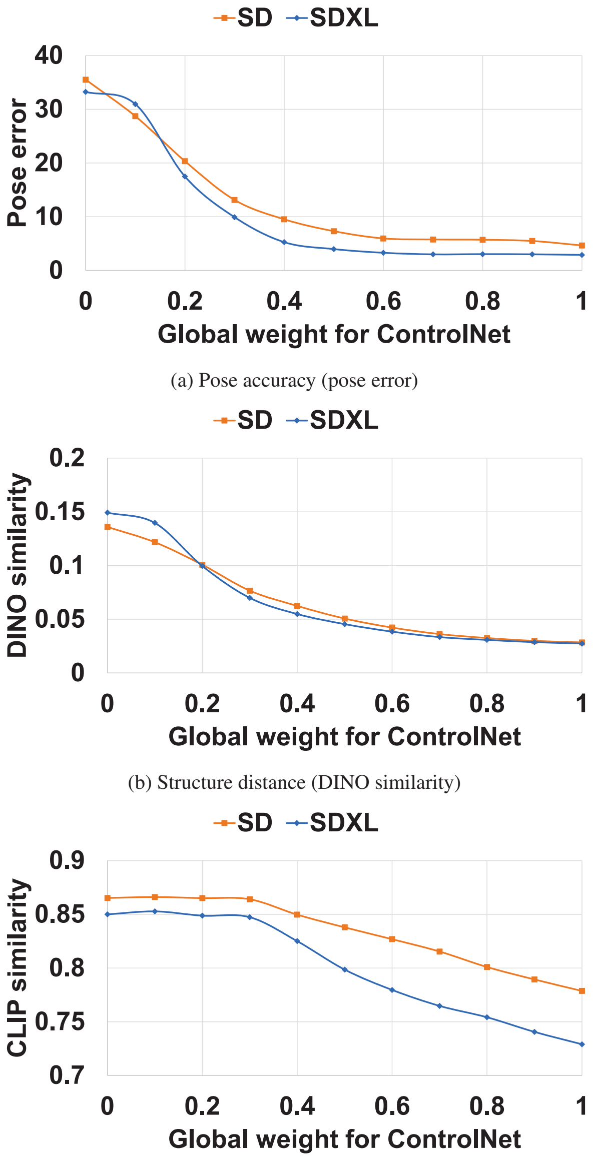
In our approach, we leverage both positive and negative text prompts, either in conjunction with image prompts or independently. For the FFHQ dataset discussed in Sec. B, the positive prompt is ”the best quality, detailed, and professional photograph of a human with muted color pupils.”, whereas the negative prompt is ”grayscale, bad anatomy, bad hands, cropped, worst quality”. Along with the images from FFHQ, to investigate the behavior of ControlNet in Sec. 4 of the main paper, we append ”side view” or ”frontal view” to the text prompt based on the target rotation poses. For animals, the positive prompt is ”the best quality, detailed, and professional photograph of a *category*,” while the negative prompt is ”vivid pupils, grayscale, bad anatomy, cropped, worst quality”. In this context, *category* refers to either a specific category depicted in the figure or wild animal for AFHQWild dataset.

## E. Global weight for ControlNet features

In ControlNet, the features from the ControlNet encoder $E_C$ are scaled by the *global weight* $\alpha$ and then added to the corresponding features of the encoder $E$ before being passed to the decoder $D$. We analyze the impact of the global weight $\alpha$ on the pose accuracy, shape incorporation, and prompt alignment of the generated images. Specifically, shape incorporation is evaluated using a structural distance based on DINO similarity [18] to assess the alignment between the shape of the conditions and the generated images. Additionally, we evaluate pose accuracy and prompt alignment as described in Sec. C. Furthermore, we conduct the analysis of ControlNet using two diffusion models: SD and SDXL. To isolate the effects of the models and exclude the impact of conditions, we use depth conditions estimated from ZoeDepth [1] for both models. The datasets used in this experiment are identical to those described in Sec. 4.1.

Initially, we investigate the impact of the global weight on *pose accuracy* and *shape incorporation* between the condition and the generated images. As shown in Fig. Aa, ControlNets based on SD and SDXL uniformly reflect the pose of the condition once their global weights reach 0.6 and 0.5, respectively, with the pose remaining unchanged beyond these values. In contrast, as demonstrated in Fig. Ab, the structure distance based on DINO similarity gradually decreases even when the global weights exceed these thresholds. From this, we conclude that ControlNets reflect the pose of the condition when the global weight is around 0.5 to 0.6, and applying higher global weights primarily affects the reflection of the shape of the condition. However, while adjusting the global weight of ControlNet helps reflect the pose of the given condition, as noted in previous studies [2, 13], relying solely on the global weight is insufficient to faithfully capture the pose. Therefore, we propose SnP to generate images that accurately reflect the pose of the conditions while faithfully incorporating the prompt, including its shape.

Additionally, we assess the impact of ControlNet's global weight on *prompt alignment*, measured using CLIP cosine similarity. As shown in Fig. Ac, it is evident that, with the same global weight, SDXL-based ControlNet sig-

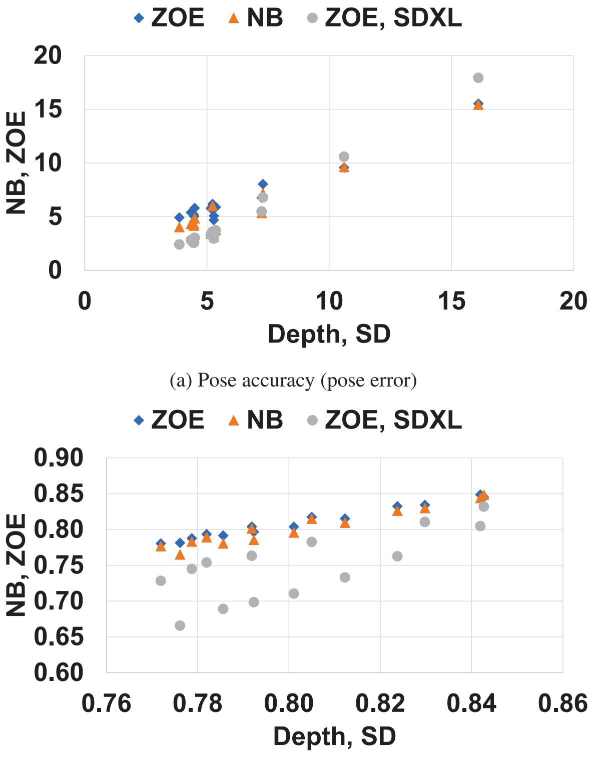


(a) Pose accuracy (pose error)

(b) Structure distance (DINO similarity)

(c) Prompt alignment (CLIP cosine similarity)

Figure A. Comparison of ControlNets based on SD and SDXL in terms of pose accuracy, shape incorporation, and prompt alignment as a function of the global weight. To evaluate these aspects between given conditions and generated images, we measure pose accuracy using pose error, shape incorporation using structural distance based on DINO similarity, and prompt alignment using CLIP cosine similarity between the given prompts and generated images. (a) Pose accuracy remains consistent with the condition once the global weight exceeds 0.5 for SD and 0.6 for SDXL, while (b) the shape continues to be reflected beyond these thresholds. (c) Additionally, prompt alignment deteriorates significantly in SDXL-based ControlNet compared to SD-based ControlNet as the global weight increases.



(a) Pose accuracy (pose error)

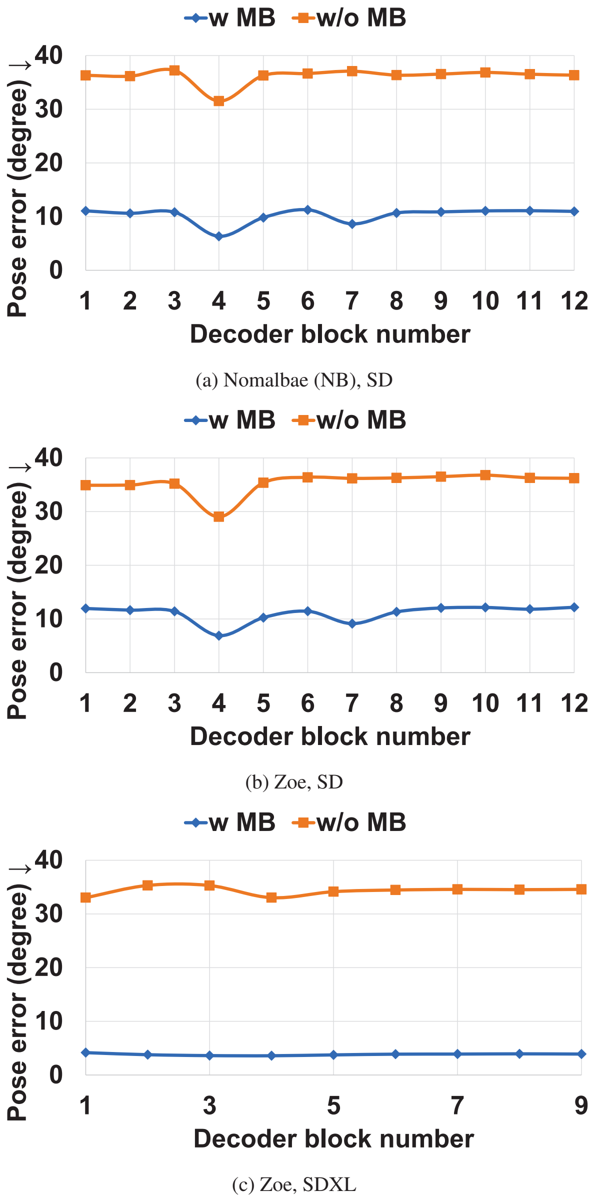
(b) Prompt alignment (CLIP cosine similarity)

Figure B. Additional analysis of the impact of ControlNet features generated from the negative prompt on the pose (a) and prompt (b) alignment of the generated images, across two additional conditions (NB, Zoe) and SDXL. The results show a positive correlation with the findings in Fig. 2, independent of the condition and diffusion model.

nificantly underperforms in prompt alignment compared to SD-based ControlNet. Therefore, to ensure comparable prompt alignment in SDXL-based ControlNet, we set its global weight to 0.6, as opposed to 1.0 in SD-based ControlNet. We note that this value is consistently applied across all SDXL experiments for both ControlNet and SnP in Sec. 5.4 of the main paper.

## F. Analysis of ControlNet According to the Diffusion Model and Condition

In this section, we conduct further analysis of the findings in Sec. 4.1 with different conditions and diffusion models. Specifically, since we conduct the previous analyses using depth conditions rendered from a 3D mesh (Sec. 4.1), we extend this analysis by incorporating two additional conditions that also contain 3D spatial information: 1) normalbae (NB), and 2) estimated depth (Zoe) from ZoeDepth [1]. Additionally, we perform the same analysis using SDXL-based ControlNet conditioned on Zoe to examine the differ-



(a) Nomalbae (NB), SD

(b) Zoe, SD

(c) Zoe, SDXL

Figure C. Additional analysis of the impact of ControlNet features passed to each decoder block on the pose accuracy of the generated images, across additional conditions (a, b) and SDXL (c). (a) and (b) use ControlNet based on SD. The experimental results indicate that the blocks influencing the pose of the generated images are determined by the diffusion model rather than the condition. Although the specific blocks that influence the pose differ between models, a commonality is that only certain blocks affect the pose of the generated images.

ences between diffusion models.

First, we conduct the same analyses using two different conditions for ControlNet, comparing the results of ControlNet conditioned on the rendered depth (Fig. 2 and Fig. 5) with those conditioned on NB and Zoe. Consequently, all outcomes conditioned on NB and Zoe demonstrate a positive correlation with the rendered depth conditions results across all analyses (Fig. B, Fig. Ca, and Fig. Cb). Especially, as depicted in Fig. Ca and Fig. Cb, the analyses of ControlNet features passed to each decoder block are consistent with the results obtained using rendered depth condition (Fig. 5). In the analyses using three conditions, the middle block (MB) has the most significant impact on pose, followed by the feature from the encoder block corresponding to the fourth decoder block. Based on these results, we conclude that SnP can be applied regardless of conditions.

Also, we compare the results of the same analyses between ControlNets based on SD and SDXL. As shown in Fig. B, the analyses of features obtained from the negative prompt in SDXL-based ControlNet show positive correlations with the results of SD-based ControlNet (Fig. 2). However, as shown in Fig. Bb, SDXL-based ControlNet exhibits lower prompt alignment compared to SD-based ControlNet, as the ControlNet features hinder prompt alignment more in SDXL-based ControlNet than in SD-based ControlNet (Fig. Ac).

The difference between the two models is highlighted in the analysis of ControlNet features passed to each decoder block, as shown in Fig. Cc. In SD-based ControlNet, two blocks influence the pose of the generated images (Fig. Cb), whereas in SDXL-based ControlNet, only the middle block (MB) affects the pose. This difference arises from structural disparities between the two models. Specifically, in SD, the roles of the middle and fourth blocks are distinct due to upsampling and three intermediate blocks. In contrast, SDXL lacks these intermediate blocks, resulting in no differentiation between the middle block and the first decoder block (which corresponds to the fourth block in SD). Despite the differences in the specific blocks that influence the pose of the generated images in each model, a commonality remains: only certain blocks affect the pose. Through this analysis, we identify structural differences between the models and apply SnP to the blocks influencing the pose, allowing the models to generate images that reflect both the pose of the condition and the content of the prompt, regardless of the diffusion model used.

## G. Additional Qualitative Results

In this section, we aim to demonstrate the superiority of our method by showcasing visual results with new objects or poses, extending from the results presented in the main paper. We validate the effectiveness of our approach by comparing it with depth-conditional ControlNet, as we use it in a training-free manner. First, in Fig. D, we generate images that reflect various text prompts and the poses of depth conditions extracted from five different objects: a car, a house, a bag, a teddy bear, and chairs. From the depth condition obtained from the car image, we generate various types of cars, while from the depth condition extracted from the house image, we produce images of diverse architectural structures. Images of various types of bags and dolls are generated from depth conditions extracted from bag and teddy bear images, respectively. Furthermore, to demonstrate the effectiveness of our method on multi-object images, we extract depth conditions from images containing two chairs and generate images of various types of chairs guided by text prompts. Comparing the results of our model and ControlNet, our method generates images that reflect the text prompt while maintaining the pose of the given depth condition. In contrast, ControlNet tends to produce objects that follow the shape of the depth condition rather than the text prompt. This tendency of ControlNet leads to the problem of images generated from the same depth condition having consistent shapes regardless of the text prompt, as already shown in the main paper.

Furthermore, we conduct a qualitative comparison between SnP and ControlNet using image prompts (Fig. E) by generating images of wild animals and humans. Even in this case, the same tendency persists. In more detail, ControlNet demonstrates a tendency to faithfully integrate the provided depth condition, which leads to the emergence of artifacts. For instance, as illustrated in the third column of Fig. Ea, it generates images of a tiger with ears resembling those of a fox, as seen across all results in Fig. Ea. Particularly noteworthy is Fig. Eb, where, despite the absence of glasses and differences in hairstyle in the image prompt, these elements appear in the generated images due to the influence of the depth condition, especially in the third, sixth, and seventh columns. These occurrences are more pronounced in ControlNet, which is heavily influenced by the depth condition, contrasting with SnP, which reduces the impact of the depth condition on the shape.

In Sec. 5.2 of the main paper, we demonstrate that our model can generate images with shapes that are independent of the depth condition but dependent on the prompts, unlike structure-guided image generation models. Taking it a step further, we present additional results in Fig. F to demonstrate that our method can generate images based on the given prompt with diverse shapes, while preserving the pose of the depth condition, even when the depth condition and the prompt belong to different categories (*e.g.*, vehicle). As evident from the results, our method generates various objects with diverse shapes from a single depth condition obtained from a bag image, in contrast to the results of depth-conditional ControlNet.

Also, we generate diverse images under identical conditions, using sampled images from FFHQ for the image prompt and sampled depth conditions from PoseH for pose control. As shown in Fig. G, images generated from different image prompts under the same depth condition exhibit the same pose but different content, aligned with the image prompts. Additionally, images generated from the same image prompt but different depth conditions show the alignment of poses with different depth conditions while maintaining the same content. We note that the same latent is used to generate images in each column.

## References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2, 3

[2] Shariq Farooq Bhat, Niloy J Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. *arXiv preprint arXiv:2312.03079*, 2023. 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 2

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020. 1

[6] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1

[8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 1

[9] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *ICIP*, pages 2496–2500. IEEE, 2022. 1, 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 2

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[12] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1, 6

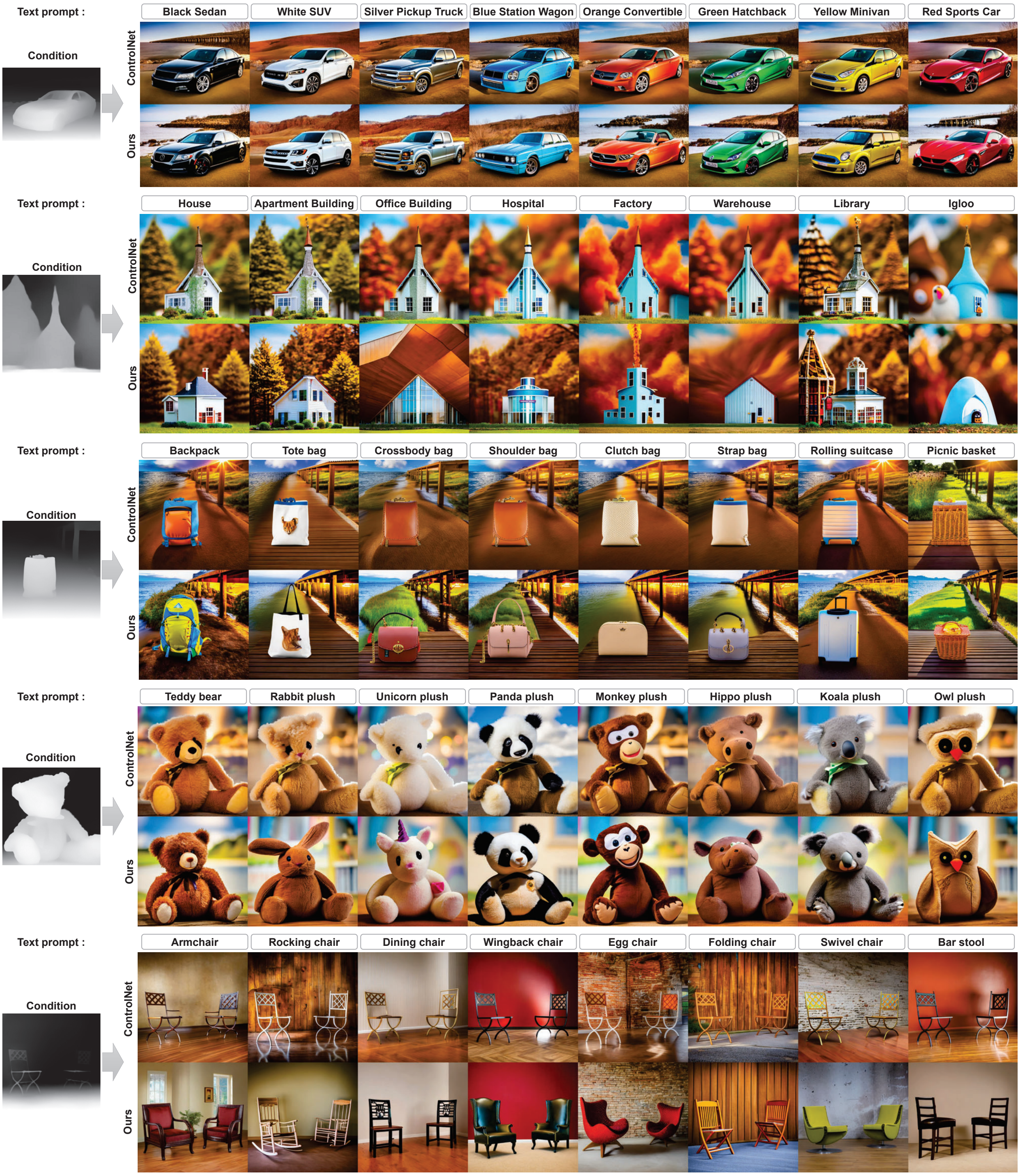


Figure D. Images of various objects generated from our method and depth-conditional ControlNet using diverse text prompts and depth conditions. Images for depth estimation of the house, bag, teddy bear, and chairs are obtained from the internet, while images of the car are sourced from CompCars [12]. Images generated by ControlNet are shape-dependent on the depth conditions, resulting in shared shapes across images generated from the same depth condition. In contrast, images produced by our model primarily reflect the provided prompts, including shape, rather than the depth conditions.

[13] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. *arXiv preprint arXiv:2404.06451*, 2024. 1, 2

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

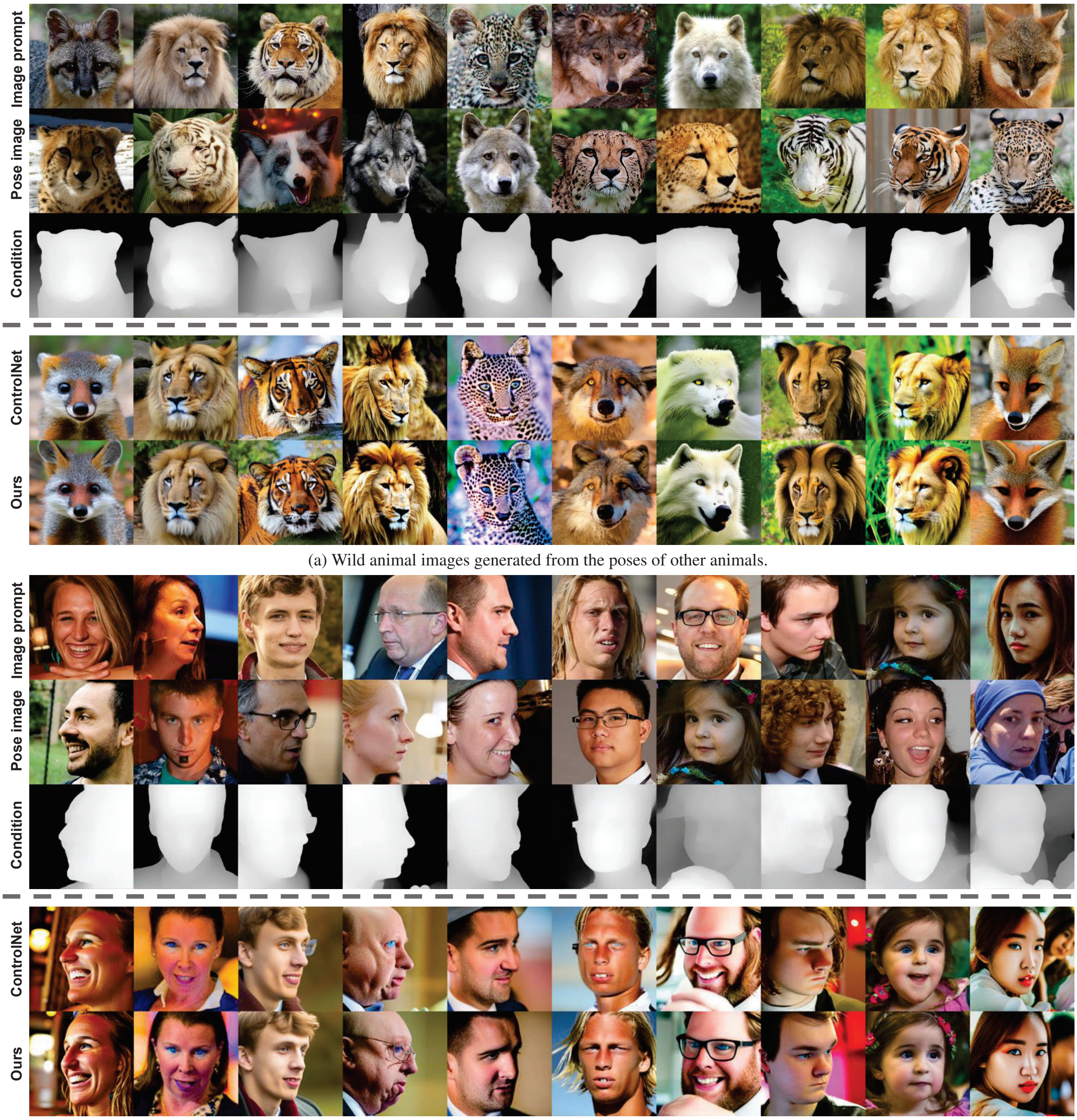
[16] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 2

[17] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[18] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 2

[19] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1



(a) Wild animal images generated from the poses of other animals.

(b) Human face images generated using the poses of other human faces.

Figure E. Qualitative results conditioned on the estimated depth maps (the third row) from pose images. The fourth and fifth rows show the results of ControlNet and our method, respectively. Images generated from ControlNet exhibit artifacts (*e.g.*, ears, hat, glasses *etc.*) to reflect the shape of the depth condition, whereas images generated from our method more accurately reflect the prompts, thus avoiding such artifacts.

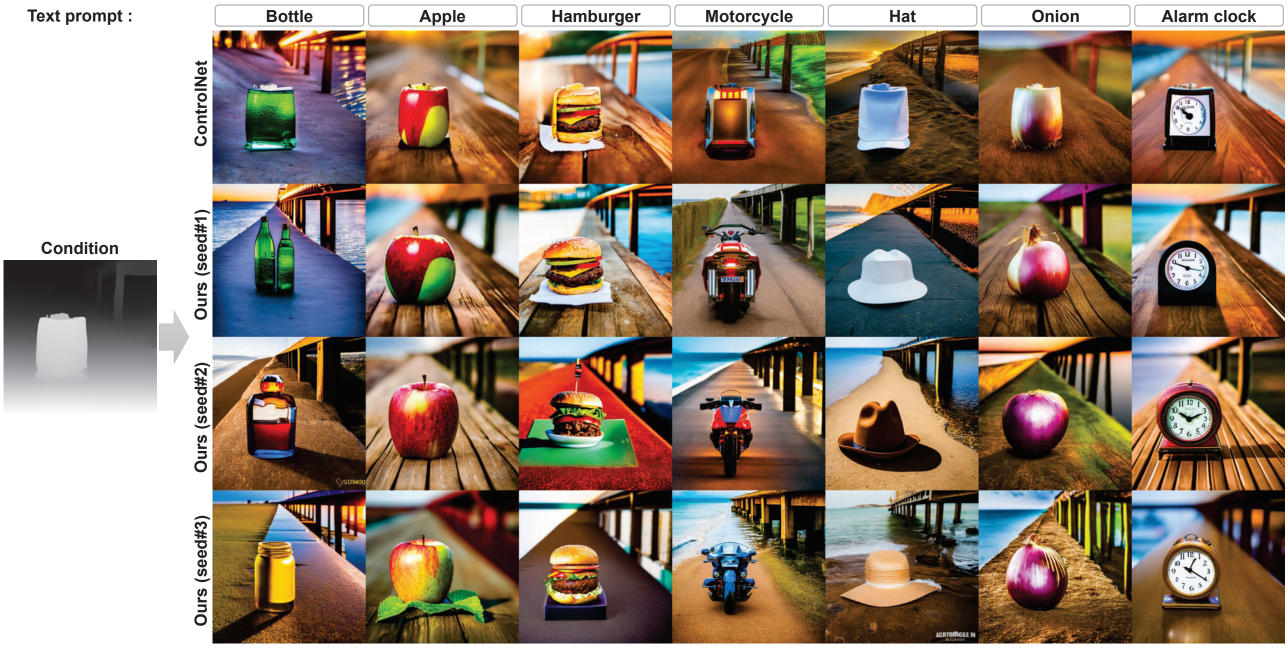


Figure F. Images of various objects generated from our method (using three different seeds) and depth-conditional ControlNet utilizing the depth condition estimated from *an entirely different object* (bag). Note that all images in each row are generated using the same latent. Also, the images in the first and second rows are generated using the same latent. Despite generating images from the depth condition extracted from an entirely different object, unlike ControlNet, SnP produces images with diverse shapes reflecting the prompt, which is also evident in images generated from different seeds.

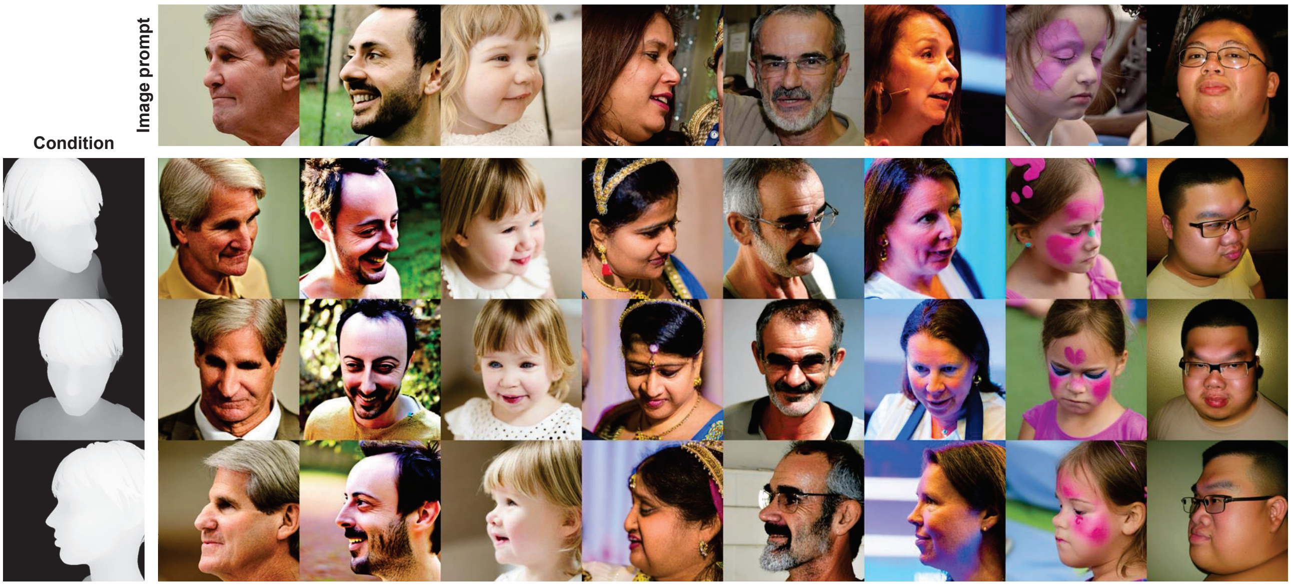


Figure G. Diverse image generation under identical conditions: depth conditions and image prompts. The images in each column are generated using the same latent and image prompt, while the images in each row are generated using the same depth condition to control the pose.