

666 A Limitations and Future Research

667 While UFC demonstrates strong few-shot performance in spatially-conditioned image generation with
 668 T2I diffusion models, several limitations remain for future work. First, our method is not designed for
 669 broader conditional generation tasks such as image manipulation (*e.g.*, style transfer, editing), inverse
 670 problems (*e.g.*, colorization, deblurring, inpainting), or spatially misaligned tasks like subject-driven
 671 generation, which often require preserving the appearance of the condition image. Extending the
 672 framework to handle such tasks is a promising direction for future research. Second, we have not
 673 explored the composition of conditions—i.e., using multiple conditions jointly to guide generation.
 674 Lastly, our approach relies on fine-tuning with a small set of annotated examples for each new task,
 675 unlike Large Language Models (LLMs), which have in-context learning capabilities, enabling them
 676 to adapt to new tasks from a few examples without fine-tuning. Developing similar capabilities for
 677 conditional image generation remains an open and promising challenge.

678 B Implementation Details

679 B.1 Settings

- 680 • **Checkpoint:** We use the Stable Diffusion v1.5 checkpoint available from the HuggingFace [7].
- 681 • **Hyperparameters:** We train using the AdamW optimizer [27] with a learning rate of 1×10^{-5}
 682 and a weight decay of 0.01. For each training batch, we randomly select two tasks per batch, each
 683 accompanied by a support set of three example pairs sampled for its query condition.
- 684 • **Spatial condition representation:** Following ControlNet [48], we represent all conditioning inputs
 685 as RGB images with a resolution of 512×512.
- 686 • **Support image-label pairs used for evaluation:** In Section G we present the five support image-
 687 label pairs used for experiments on six tasks. Specifically, Figure 13 presents the pairs used for
 688 Canny, Depth, HED, and Normal, while Figure 14 presents those used for Pose and DensePose.

689 B.2 Dataset

- 690 • **Training data:** We randomly sample a subset of data from the LAION dataset [38]. To identify
 691 images containing humans, we use the YOLO11x model [17]. Based on this filtering, we construct
 692 a balanced dataset consisting of 150K images with humans and 150K images without humans.
- 693 • **Evaluation data:** For the Canny, HED, Depth, and Normal tasks, we evaluate our model on 5,000
 694 images from the COCO2017 validation set [25]. For the Pose task, evaluation is limited to images
 695 where humans are successfully detected by the Openpose model [6]. Similarly, for the Densepose
 696 task, we only evaluate our model with images where the Densepose model [11] detects a human
 697 subject.

698 B.3 Matching Module Implementation

699 We implement our matching module using the multi-head attention mechanism [42]. To incorporate
 700 denoising timestep t , we adopt *adaptive normalization* [31], modulating the output of the matching
 701 module using the denoising timestep. Specifically, we use the embedding of the query condition
 702 $g_\tau(y_\tau^q)$, the support conditions $\{g_\tau(y_\tau^i)\}_{i=1}^N$, and the support images $\{f(x^i)\}_{i=1}^N$ as the query, key,
 703 and value inputs, respectively. Let $Q \in \mathbb{R}^{M \times d}$, $K, V \in \mathbb{R}^{(N \cdot M) \times d}$, and the timestep embedding be
 704 denoted as $t_{\text{emb}} \in \mathbb{R}^{d_t}$, our matching mechanism operates as below:

$$Q = \text{LayerNorm}(Q), \quad K = \text{LayerNorm}(K), \quad (6)$$

$$(\alpha, \beta, \gamma) = \text{Linear}(t_{\text{emb}}), \quad V = \text{LayerNorm}(V) \cdot (1 + \alpha) + \beta \quad (7)$$

$$O = \text{Concat}_{i=1}^H \left(\text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \right) W^O \quad (8)$$

707 where H is the number of attention heads, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_{\text{head}}}$, and $W^O \in \mathbb{R}^{H \cdot d_{\text{head}} \times d}$.

708 The final output is calculated with residual connection as below:

$$O = O + \gamma \cdot \text{act}(O) \quad (9)$$

709 where act denotes a non-linear activation function.

710 We apply the matching module at each layer of the 12 down-sampling blocks and the mid-block in
 711 the UNet backbone of the diffusion model. Each attention module uses 8 heads.

712 C More Results on Analysis

713 **Attention map visualization** We present an example of attention maps in Figure 5. For each task,
 714 the selected query patch (highlighted by a white box in the Query column) can attend to relevant
 715 support patches, rather than unrelated regions such as background areas. For instance, in the Canny
 716 and HED tasks (first two rows), the query patches focus on support regions that preserve similar edge
 717 structures. On the other hand, for the Pose and Densepose tasks (last two rows), the query patches
 718 attend to regions related to human body parts.



Figure 5: Attention Maps of UFC (30-shot) from the 7th layer and a selected head for each task. Support normal maps are converted to grayscale to enhance the visibility of attention regions. Query patches attend primarily to the most relevant support patches.

719 **Qualitative results with two variants** Figure 6 presents a qualitative comparison of our method
720 with its two variants: (1) UFC w/o Matching and (2) UFC w/o finetuning. Both variants exhibit
721 degraded image generation in terms of controllability, as they often only partially adhere to the spatial
722 conditions. For instance, in the 2nd row with HED condition, the generated images from UFC w/o
723 Matching and UFC w/o Fine-tuning capture the cat’s head structure but fail to adhere to the condition
724 for the cat’s legs and body. In contrast, our full method, which include both matching and fine-tuning,
725 consistently follows all given conditions.

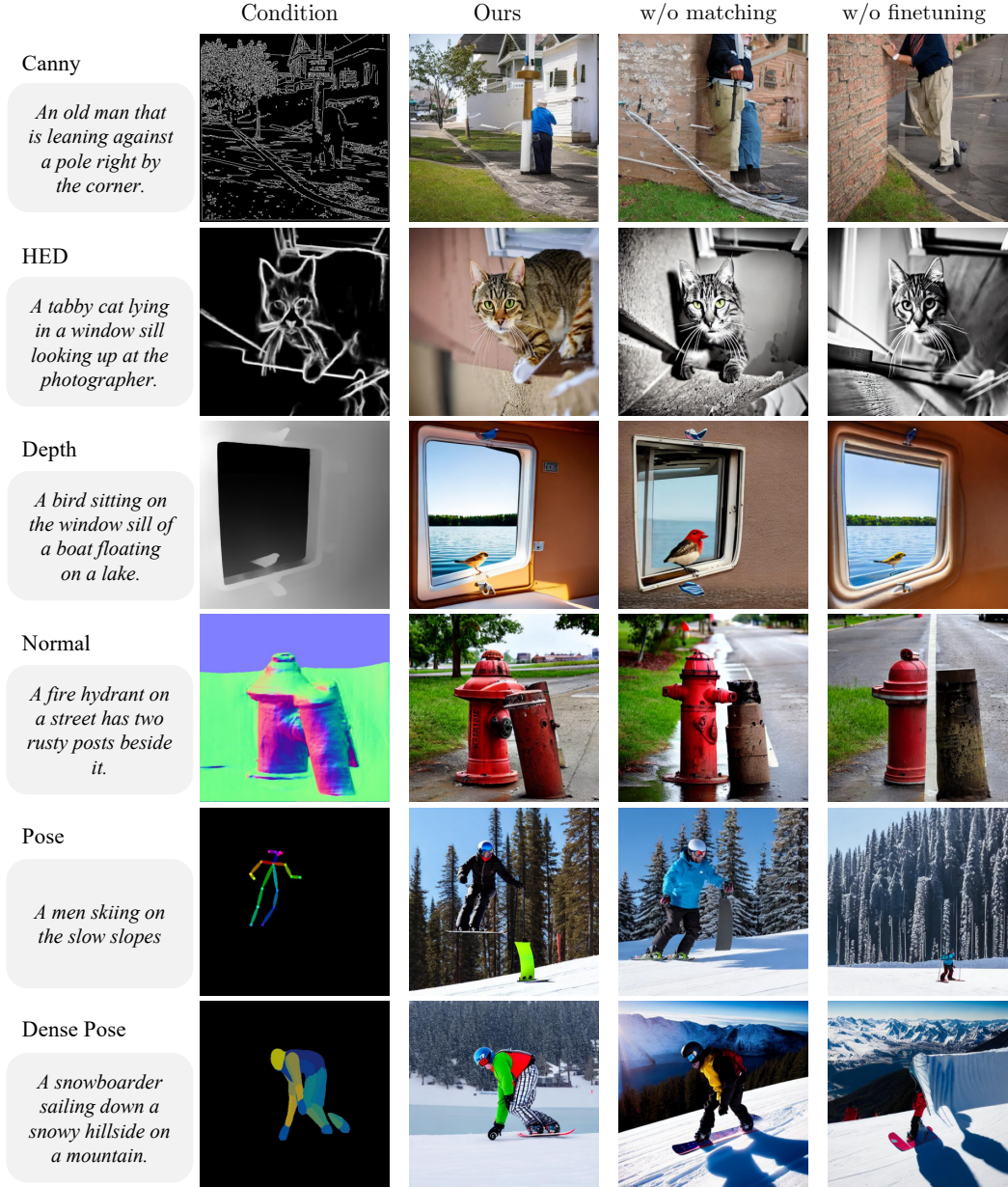


Figure 6: Qualitative comparison results of our method and its two variants: (1) UFC w/o matching and (2) UFC w/o fine-tuning.

726 **Results on FID with different number of shots** Figure 7 shows the FID results obtained by
727 fine-tuning our method (UFC) using different numbers of support data (i.e., shots). The results show
728 that our method maintains FID scores across various shots. Specifically, for Normal, Depth, Canny,
729 and HED tasks, our method preserves image quality with FID changes remaining within 1. For Pose,

the FID difference is within 4.5, and for DensePose, it is within 1.7. Both changes are relatively small given the FID scores of each tasks, and using more support data often slightly improves FID.

Combined with the controllability results measured using different number of shots (Figure 4), the results confirm that our method improves controllability as the number of support data increases, while maintaining image quality.

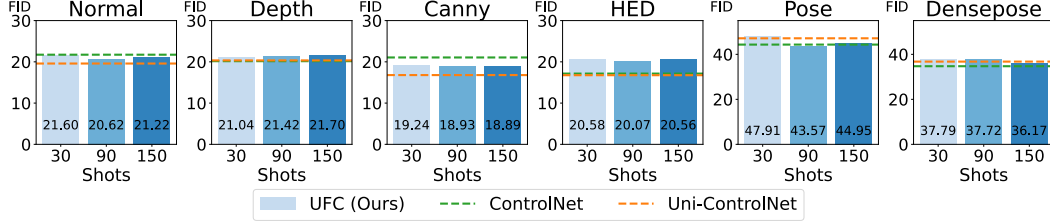


Figure 7: FID score over varying support set sizes (shots).

D Results with DiT

D.1 Implementation Details

Model We extend our framework to the DiT architecture [30], using Stable Diffusion v3 [8] as a test case, initialized from the v3.5-medium checkpoint [1] on Hugging Face. Due to our computational resource constrains, it is impractical to full fine-tune the label encoder with DiT backbone. To address this, we freeze and share the weights of the diffusion model, label encoder, and image encoder, training only the matching module and the bias parameters for the label encoder. Despite this minimal setting, we demonstrate a proof of concept for DiT-based architectures, which are behind the recent advanced T2I models [8, 21].

Control feature injection Unlike the U-Net backbone, the DiT architecture does not have skip connections. Therefore, we inject the output of the matching module directly into the hidden representations at each layer of the DiT backbone. Specifically, we apply the matching module to 12 of the 24 transformer layers—namely, the even-numbered layers (0, 2, 4, ...), balancing memory efficiency and performance.

Hyper-parameters We follow the same optimizer settings, training dataset, and evaluation protocol described in Section 5.1. For inference, we use the flow-matching Euler scheduler introduced in Stable Diffusion 3 [8], with a classifier-free guidance (CFG) scale of 5.0, 28 generation steps, and a fixed random seed of 42.

D.2 Experimental Results

Table 4: Quantitative evaluation of UFC with different backbones in 30-shot setting.

Backbone	Canny		HED		Depth		Normal		Pose		Densepose	
	SSIM↑	FID	SSIM↑	FID	MSE↓	FID	MAE↓	FID	AP ⁵⁰ ↑	FID	mIoU↑	FID
Ours (UNet)	0.3239	19.24	0.5121	20.58	94.38	21.04	15.09	21.60	0.229	47.91	0.434	37.79
Ours (DiT)	0.3757	20.91	0.5703	20.77	92.62	23.28	15.43	23.55	0.107	52.57	0.454	38.36

We report quantitative results using the DiT backbone in Table 4. While the DiT-based model outperforms the U-Net backbone on 5 out of 6 tasks in terms of controllability (excluding Pose), its image quality is not as good. A similar trend is observed in OmniControl [40], where DiT shows lower visual quality for spatial control tasks. For the Pose task, DiT performs worse than its U-Net counterpart, which we attribute to the spatial sparsity of pose conditions and the limitations of our minimal implementation—particularly the lack of fine-tuning for the label encoder. Further analysis, such as examining attention maps for the Pose task, is needed to better understand this limitation. We leave this investigation for future work. Qualitative examples using the DiT backbone are shown in Figure 8.

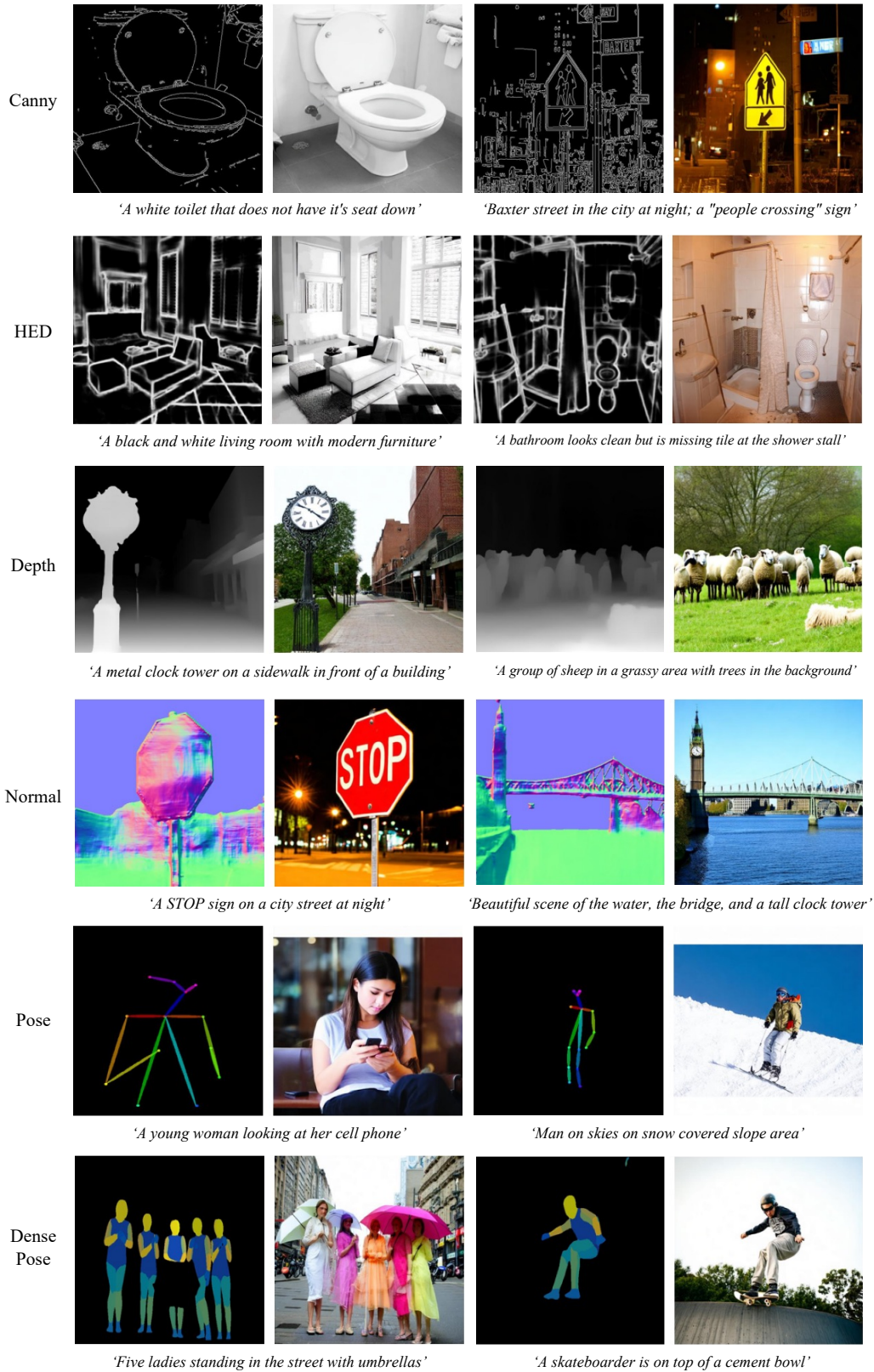


Figure 8: Generated images from UFC with DiT backbone in 30-shot setting.

E Comparison with Training-free Baselines

We propose a few-shot framework for adapting to new spatial conditions in T2I diffusion models. A natural question arises: *why use a few-shot approach when training-free methods exist [2, 47, 28]?* To illustrate the advantages of our method, we compare it with FreeControl [28], a state-of-the-art training-free approach capable of handling diverse spatial conditions without relying on additional pretrained networks. FreeControl controls image structure by constructing a PCA basis from object features, projecting both the condition and noisy image feature maps onto this basis, and minimizing an energy function that encourages alignment between the two projections during generation.

Evaluation protocol We follow FreeControl’s original evaluation protocol and assess controllability on 30 images from the ImageNet-R-TI2I dataset [41], which includes 10 object categories with three captions each. FreeControl has limited flexibility in generating diverse object categories, as it requires constructing a separate PCA basis for each. This makes large-scale evaluation on 5,000 images from the COCO2017 validation set [25] impractical, since each prompt must be individually inspected to identify objects and build corresponding PCA bases.

As the evaluation dataset of FreeControl excludes human subjects, we compare UFC (30-shot) against FreeControl on four tasks: Canny, HED, Depth, and Normal. Due to the limited number of images, we do not report FID [12] or Inception Score [37], as they would not provide reliable estimates of image quality.

Result comparison Table 5 shows that UFC (30-shot) significantly outperforms FreeControl in controllability across all four tasks. Moreover, on a single NVIDIA RTX 3090 GPU, FreeControl takes approximately 100 times longer per image to generate compared to UFC when generating 30 images for evaluation. This overhead stems from the need to construct a PCA basis for each new object category and perform 200 denoising steps for latent optimization. In contrast, UFC, after fine-tuning, handles diverse object categories flexibly with only 50 denoising steps.

Table 5: Controllability scores and average generation time of FreeControl [28] and UFC (Ours, 30-shot) on generating 30 images from ImageNet-R-TI2I [41].

Method	Canny SSIM (↑)	HED SSIM (↑)	Depth MSE (↓)	Normal MAE (↓)	Time (Second)
FreeControl [28]	0.3139	0.3821	97.36	19.68	251.3
UFC (Ours)	0.4074	0.5718	93.09	17.75	2.5

Figure 9 presents qualitative comparisons. UFC accurately follows the spatial condition, while FreeControl fails on fine details—especially under the Normal condition, where multiple objects present in the condition.

F More Results

Results with more spatial conditions To further validate the generalizability of our method across diverse spatial conditions, we present additional qualitative results using single-view images derived from 3D meshes, wireframes, and point clouds. These spatial conditions are challenging to obtain from existing images, making it difficult to collect large-scale training data. We use a 3D isolated object dataset¹ (with no background) to generate spatial condition to image pairs.

Figure 10 shows results of our method, demonstrating generated images that align closely with the spatial conditions provided by 3D meshes, wireframes, and point clouds. These results confirm UFC’s effectiveness when encountered to novel spatial conditions.

More qualitative results We present more qualitative results of UFC in Figure 11 and Figure 12. All the results are generated with the support size of 30.

¹<https://huggingface.co/datasets/dylanebert/iso3d>

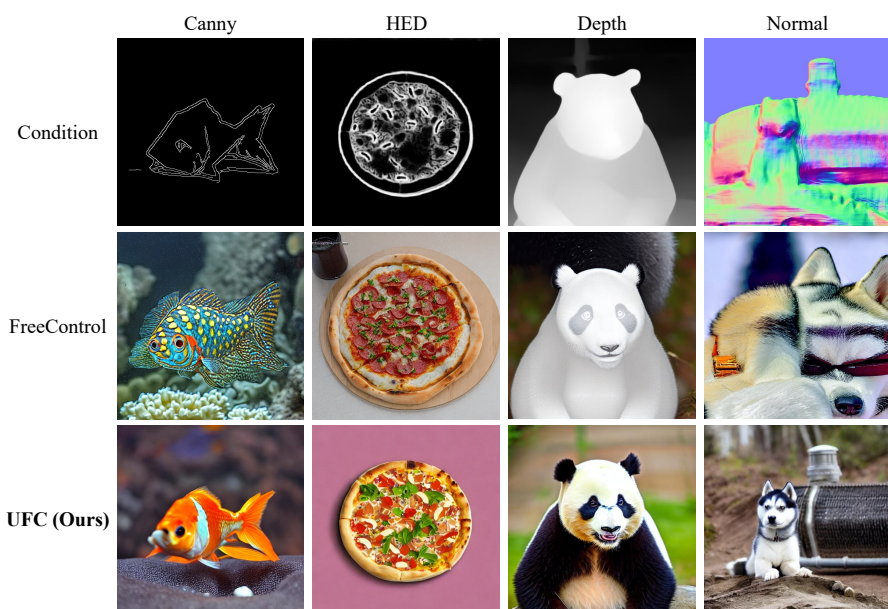


Figure 9: Qualitative comparison between UFC (30-shot) and FreeControl [47] on four control tasks

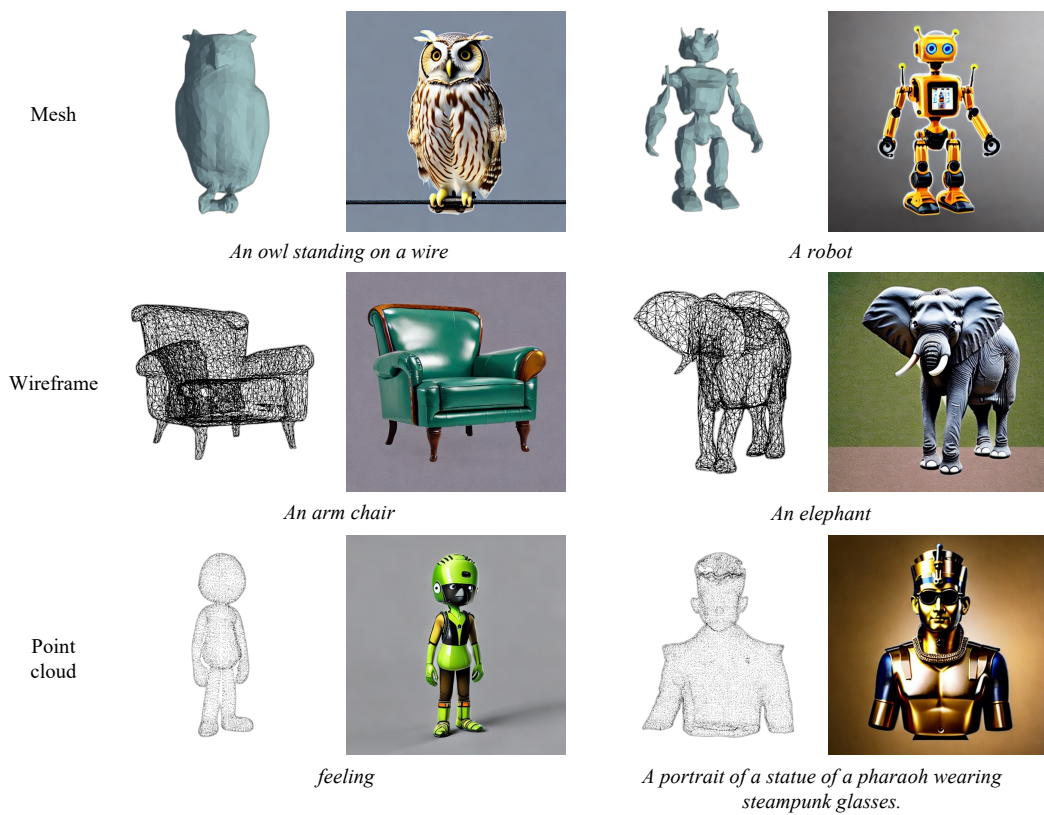


Figure 10: Generated images from UFC with spatial conditions of 3D meshes, wireframes, and point clouds in 30-shot setting.

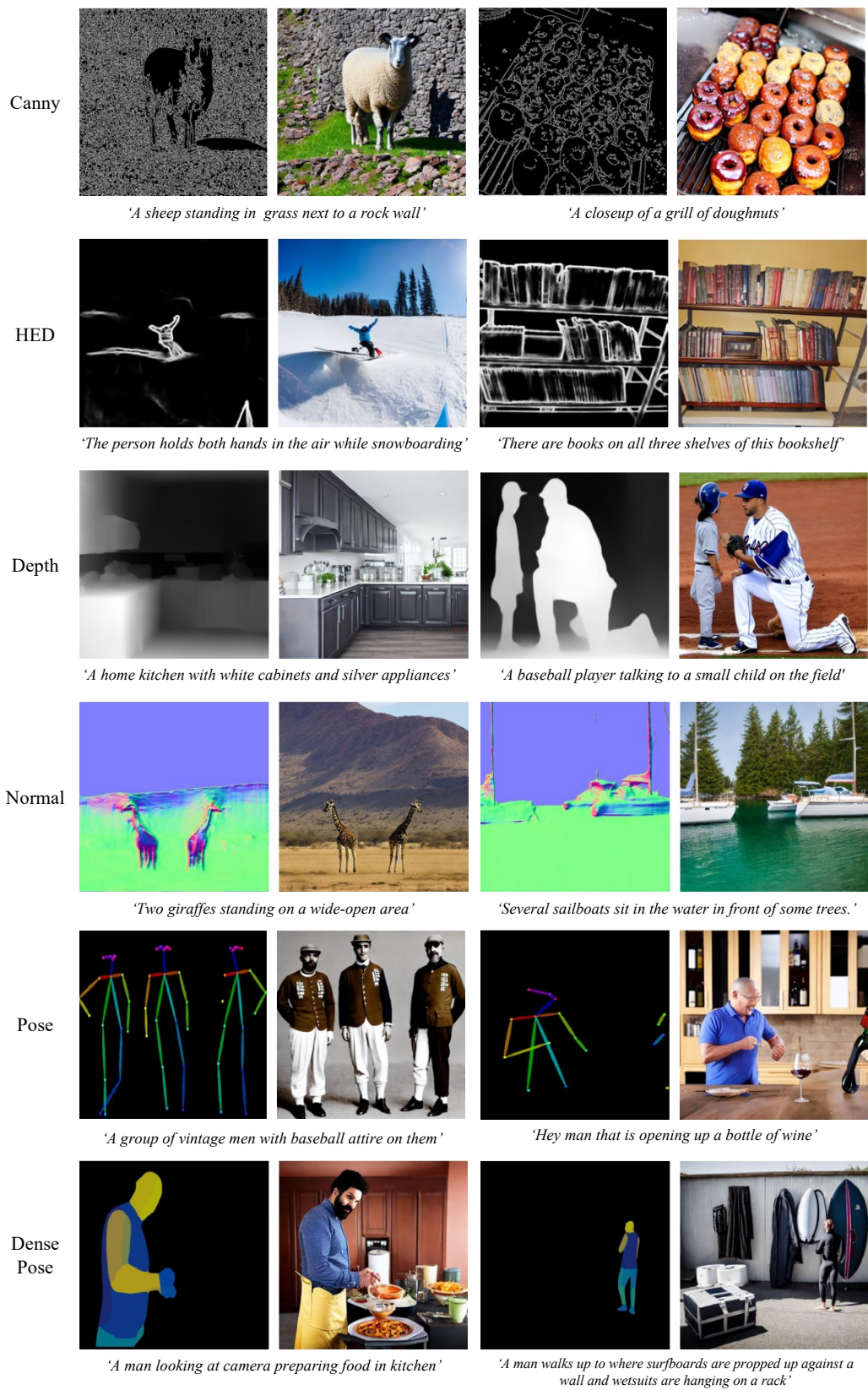


Figure 11: More qualitative results of UFC in 30-shot setting.

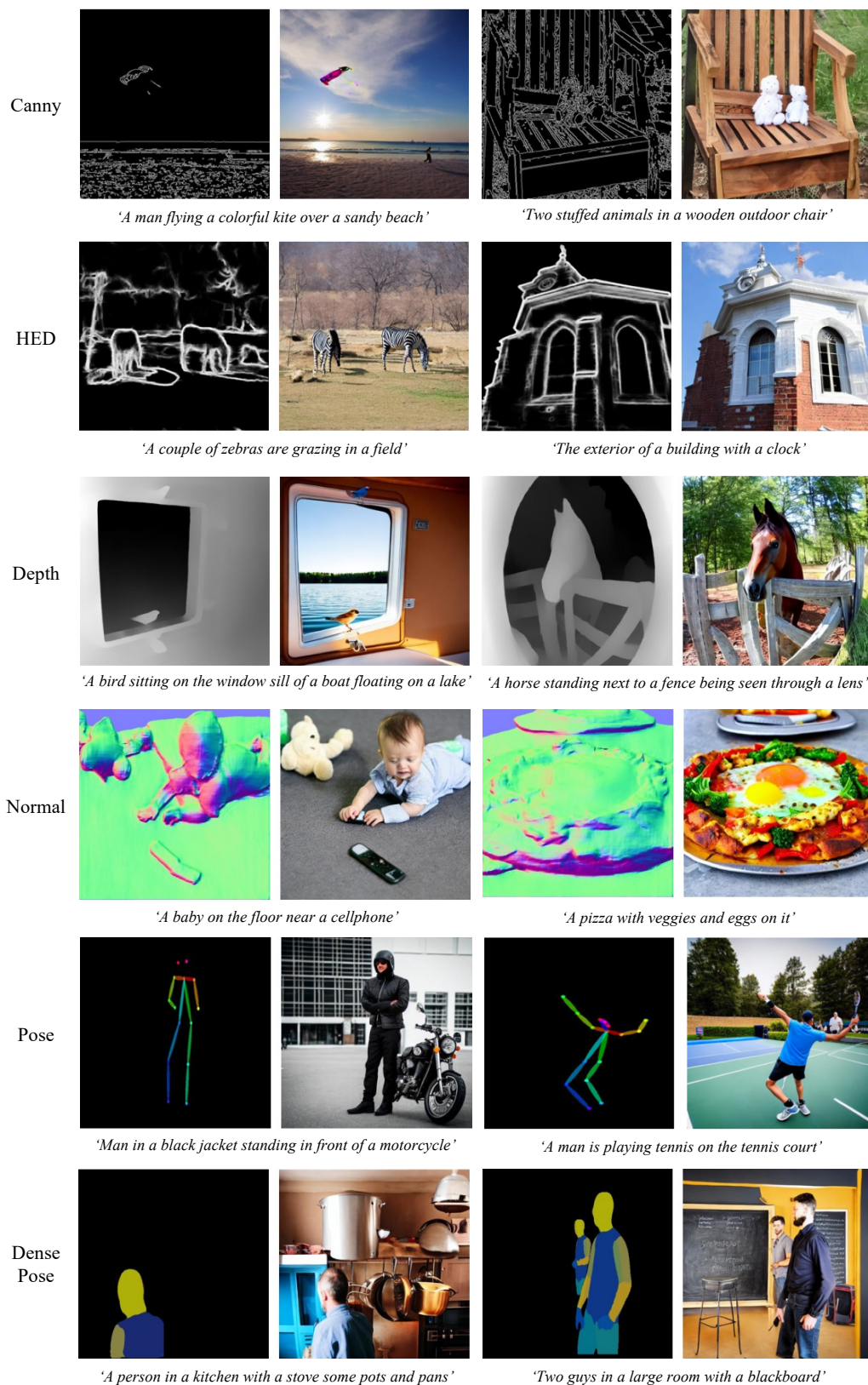


Figure 12: More qualitative results of UFC in 30-shot setting.

801 **G Support Image-Condition Pairs**

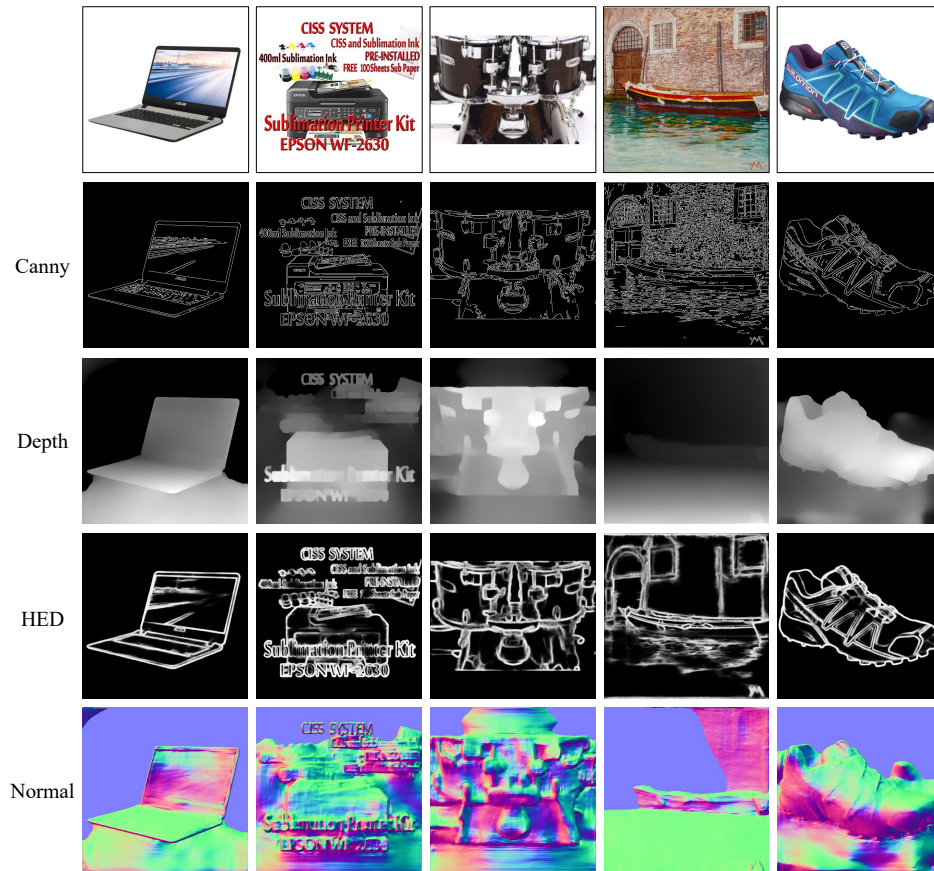


Figure 13: Five support image-label pairs used for evaluating Canny/Depth/HED/Normal tasks.

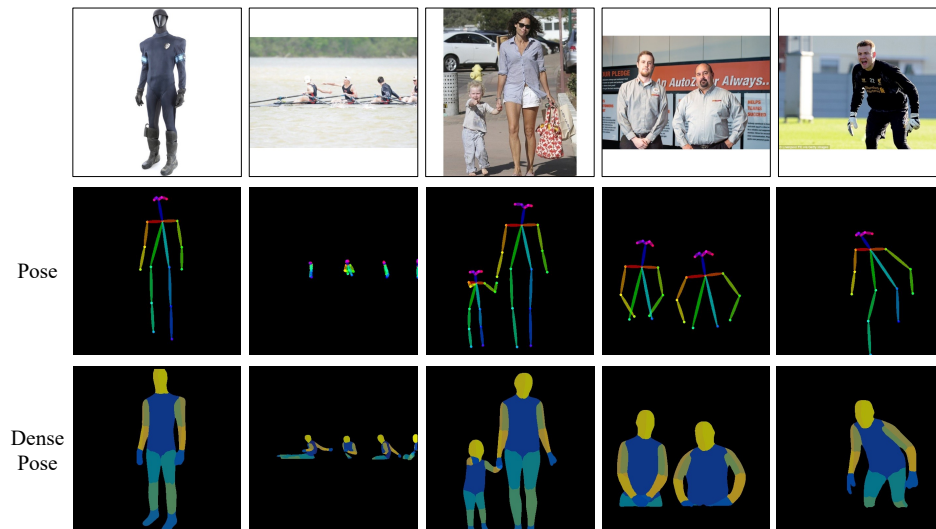


Figure 14: Five support image-label pairs used for evaluating Pose/DensePose tasks.

References

- [1] Stability AI. 2024. Stable Diffusion 3.5 Medium. <https://huggingface.co/stabilityai/stable-diffusion-3.5-medium>. Accessed: 2025-05-22.
- [2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Universal Guidance for Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 843–852.
- [3] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.
- [4] John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.
- [5] Yu Cao and Shaogang Gong. 2024. Few-shot image generation by conditional relaxing diffusion inversion. In *European Conference on Computer Vision*, pages 20–37. Springer.
- [6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [7] CompVis. Stable Diffusion v1-5. <https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>. Accessed: 2025-05-22.
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *CoRR*, abs/2403.03206.
- [9] Giorgio Giannone, Didrik Nielsen, and Ole Winther. 2022. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*.
- [10] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. 2024. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4):1–15.
- [11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306.
- [12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- [14] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- [15] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. 2023. Composer: Creative and Controllable Image Synthesis with Composable Conditions.
- [16] Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, and Yabiao Wang. 2025. Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation.
- [17] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. Ultralytics YOLO.

- [18] Donggyun Kim, Seongwoong Cho, Semin Kim, Chong Luo, and Seunghoon Hong. 2024. Chameleon: A data-efficient generalist for dense visual prediction in the wild. In *European Conference on Computer Vision*, pages 422–441. Springer.
- [19] Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, and Seunghoon Hong. 2023. Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching. In *The Eleventh International Conference on Learning Representations*.
- [20] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- [21] Black Forest Labs. 2024. FLUX.1-schnell. <https://huggingface.co/black-forest-labs/FLUX.1-schnell>. Accessed: 2025-05-22.
- [22] Lingxiao Li, Yi Zhang, and Shuhui Wang. 2023. The Euclidean Space is Evil: Hyperbolic Attribute Editing for Few-shot Image Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22714–22724.
- [23] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2024. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VII*, page 129–147, Berlin, Heidelberg. Springer-Verlag.
- [24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521.
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- [26] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *International Conference on Learning Representations*.
- [27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [28] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475.
- [29] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304.
- [30] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.
- [31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- [32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

- [33] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. 2023. UniControl: a unified diffusion model for controllable visual generation in the wild. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- [34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637.
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.
- [36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.
- [37] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.
- [38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- [39] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548.
- [40] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*.
- [41] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930.
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- [43] Zhendong Wang, Yifan Jiang, Yadong Lu, yelong shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. 2023. In-Context Learning Unlocked for Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- [44] Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1403.
- [45] Ruofeng Yang, Bo Jiang, Cheng Chen, Baoxiang Wang, Shuai Li, et al. 2024. Few-shot diffusion models escape the curse of dimensionality. *Advances in Neural Information Processing Systems*, 37:68528–68558.
- [46] Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. 2023. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems*, 36:48723–48743.
- [47] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. 2023. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184.

- 943 [48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-
944 image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer*
945 *vision*, pages 3836–3847.
- 946 [49] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and
947 Kwan-Yee K Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models.
948 *Advances in Neural Information Processing Systems*, 36:11127–11150.
- 949 [50] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. 2022. Few-shot image generation
950 with diffusion models. *arXiv preprint arXiv:2211.03264*.