Meta ControlNet: Enhancing Task Adaptation via Meta Learning

Junjie Yang[1]*, Jinze Zhao[2]*, Peihao Wang[2], Zhangyang Wang[2], Yingbin Liang[1]
[1]Ohio State University, [2]University of Texas at Austin
yang.4972@osu.edu, jz24694@utexas.edu, peihaowang@utexas.edu, atlaswang@utexas.edu,
liang.889@osu.edu
(*Equal contribution)

Diffusion-based image synthesis has attracted extensive attention recently. In particular, ControlNet that uses image-based prompts exhibits powerful capability in image tasks such as canny edge detection and generates images well aligned with these prompts. However, vanilla ControlNet generally requires extensive training of around 5000 steps to achieve a desirable control for a single task. Recent context-learning approaches have improved its adaptability, but mainly for edge-based tasks, and rely on paired examples. Thus, two important open issues are yet to be addressed to reach the full potential of ControlNet: (i) zero-shot control for certain tasks and (ii) faster adaptation for non-edge-based tasks. In this paper, we introduce a novel Meta ControlNet method, which adopts the task-agnostic meta learning technique and features a new layer freezing design. Meta ControlNet significantly reduces learning steps to attain control ability from 5000 to 1000. Further, Meta ControlNet exhibits direct zero-shot adaptability in edge-based tasks without any finetuning, and achieves control within only 100 finetuning steps in more complex non-edge tasks such as Human Pose. Our code is publicly available at https://github.com/JunjieYang97/Meta-ControlNet.

1. Introduction

Image synthesis [1–3] is a rapidly growing field in computer vision and draws significant interest from various application domains. As a key approach in this area, Generative Adversarial Networks (GANs) [3–5] employ a discriminator-generator pair, where the generator is trained to creating enhanced images via sharpening the discriminator. However, such an adversarial approach typically has difficulty to model more complex distributions. Recently, diffusion models [1, 2] have emerged as a powerful alternative, excelling in high-quality image generation. These models utilize a series of denoising autoencoders to progressively refine an image from pure Gaussian noise. Among these, a new model known as Stable Diffusion [6] has been proposed, which has better computational efficiency. Unlike traditional methods, Stable Diffusion uses latent representations for image compression, and achieves superior image quality, which includes advancements in text-to-image synthesis and unconditional image generation.

ControlNet [7] further advances image synthesis with enhanced control over image content by using conditional control as different tasks. This approach clones the encoder and middle block of Stable Diffusion, and introduces zero convolution to link with the decoders of Stable Diffusion. Such a setup allows ControlNet to accept image prompt inputs, such as canny or HED edge, and can generate images specific to certain tasks, demonstrating improved control from both image and textual inputs. However, ControlNet's capability for precise control requires extensive training. Specifically, learning to control a new task demands about 5000 steps. Recently, Prompt Diffusion was proposed in [8], which leverages in-context learning idea to enhance ControlNet's adaptability to new tasks, but requires task-specific example pairs for training.

Although ControlNet and its variants have achieved enhanced the generalizability, several critical open issues remain unresolved to reach the full power of ControlNet. **Firstly**, *zero-shot* capability

Second Conference on Parsimony and Learning (CPAL 2025).

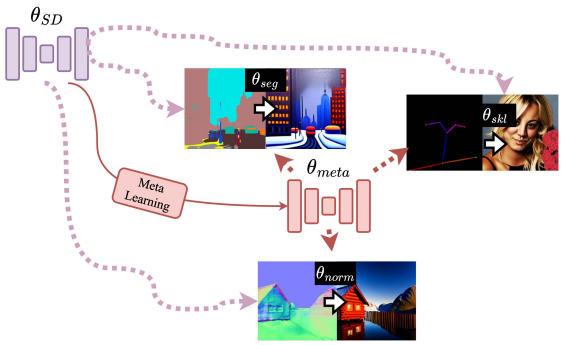

Figure 1: Trained from stable diffusion initial θ_{SD}, the meta learned initial θ_{meta} is used for various task adaptation.

of ControlNet has not yet been explored, leaving it as an open question whether it is possible to control new tasks without finetuning samples. **Secondly**, while most existing studies have focused on edge-based tasks, rapid adaptation in more complex scenarios, such as the human pose task, has not yet been achieved.

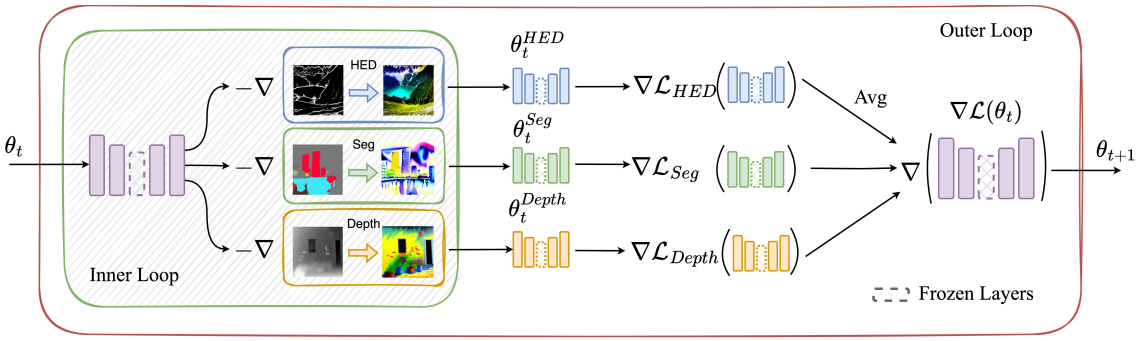

Figure 2: Meta ControlNet training pipeline. ControlNet parameter is meta updated via meta tasks (HED, Segmentation, Depth). Stable Diffusion parameters are fixed and ControlNet middle layers (Encoder Block 4 and Middle Block) are frozen during the training phase.

In this paper, we propose a novel **Meta ControlNet** method to address the aforementioned open issues. Specifically, Meta ControlNet adopts the FO-MAML [9] framework with various image condition types serving as different meta tasks. The inner-loop training of Meta ControlNet takes finetune steps separately for each task. Then the outer-loop training updates meta parameters (i.e., the model initial) based on averaged gradients over all training tasks.

(**Novel Layer Freezing Design**) Meta ControlNet features a new layer freezing design. Typically, meta learning algorithms such as ANIL [10] freezes the earlier embedding layers during the inner-loop training. As a sharp difference, Meta ControlNet freezes *latter* encoder block and the *middle* block during meta training. This idea is based on the observation that the initial encoder blocks are directly linked to the control images of control tasks. It is essential to finetune these encoder blocks for individual task. On the other hand, the middle and latter encoder blocks, which capture common and high-level information, can be retained and shared across tasks. Such a design has been proven to be critical for Meta ControlNet to exhibit desirable performance in our experiments. Note that for the meta testing phase, we recommend training all layers to achieve the best possible adaptation performance.

In the following, we highlight the superior experimental performance that Meta ControlNet achieves:

- **Fast Learning of Control Ability:** Our proposed design significantly enhances the efficiency of ControlNet's learning process. Our experiments demonstrate that Meta ControlNet acquires

control abilities within only 1000 steps, a stark improvement over the vanilla ControlNet that achieves the same ability with 5000 steps. Meanwhile, this efficiency is demonstrated across three meta training tasks, showcasing the method's versatility.

- **Zero-Shot Adaptation for Edge-based Tasks:** Meta ControlNet produces generalizable model initial, which exhibits exceptional control capabilities in zero-shot settings, especially for edge-based tasks such as the canny task. This indicates that our model can adapt to new edge-based tasks without any task-specific finetuning. This is the *first* achievement of successful zero-shot adaptation by ControlNet. Our experiments also indicate that few-shot finetuning often further enhances the fidelity of generated images.
- **Fast Adaptation in Non-edge Tasks:** For challenging tasks in few-shot contexts, our learned model initial exhibits a robust adaptation ability. For instance, it can adapt to the human pose task within merely 100 steps and excel in the more complex human pose mapping task in only 200 steps. These achievements not only surpass all existing benchmarks but also substantially reduce image sample number required for adaptation.

2. Meta ControlNet

In this section, we first propose the Meta ControlNet method, and then explain how to select and structure both training and adaptation tasks.

2.1. Algorithm Design

In this section, we propose our algorithm Meta ControlNet, which maintains the Stable Diffusion network while training its duplicates via the task-agnostic meta learning technique for obtaining adaptive model initial.

In particular, Meta ControlNet adopts three control tasks (HED, Segmentation, Depth) as the primary meta tasks. The training of Meta ControlNet takes the double-loop training framework of FO-MAML [9], as depicted in Figure 2, and is described in detail as follows.

The *inner-loop* training of Meta ControlNet takes finetune steps separately for each task. During each step t, the meta parameter θ_t of Meta ControlNet is finetuned independently for each task based on gradient descent as follows:

$$\text{(Inner Loop)} \quad \theta_t^{task} = \theta_t - \alpha \nabla \mathcal{L}_{task}(\theta_t),$$

where $task \in \{\text{HED, Seg, Depth}\}$ and α represents the step size. Note that we here update the parameter only once in the inner loop to enhance efficiency.

The *outer-loop* training first calculates the meta gradient $\nabla \mathcal{L}(\theta_t)$ as follows by taking an average of the gradients across all tasks, based on each task's finetuned parameters in the inner loop:

$$\nabla \mathcal{L}(\theta_t) = \text{Avg}_{task}(\nabla \mathcal{L}_{task}(\theta_t^{task})),$$

where "Avg" denotes the averaging operator over all task gradients. Then the meta parameter θ_t is updated using the meta gradient as follows:

$$\text{(Outer Loop)} \quad \theta_{t+1} = \theta_t - \alpha \nabla \mathcal{L}(\theta_t),$$

where α is the step size. This design guides Meta ControlNet to minimize the loss of the finetuned model for each task, and hence makes the model more responsive to updates and enables fast adaptability.

Novel Layer Freezing Design: A main novel component that Meta ControlNet features is the design of freezing layers during training, which turns out to be critical for its superior performance. Typically, meta learning algorithms such as ANIL [10] freeze the earlier embedding layers during the inner-loop training. Meta ControlNet has sharp differences in two aspects. **Firstly,** Meta ControlNet freezes *latter* encoder block and the *middle* block during meta training. This idea is based on the observation that the initial encoder blocks are directly linked to the control images of control tasks. Given that our

Figure 3: Validation set samples from each training task (HED, Depth, Segmentation) after 1000 steps of training updates.

Stable Diffusion initial was used to process Gaussian noise rather than control task-specific inputs, and different control image styles signify distinct tasks, it is essential to finetune these initial encoder blocks for individual tasks. On the other hand, the middle and latter encoder blocks, which capture common and high-level information, can be retained and shared across tasks. Therefore, during the training process, Meta ControlNet selectively freezes the Encoder Block4 and Middle block of the U-Net, and focuses on training the remaining parameters. The detailed architecture is available in Appendix. **Secondly,** unlike ANIL, where freezing occurs only in the inner loop, Meta ControlNet applies layer freezing in both inner and outer loops to achieve better efficiency by leveraging our network's initial high-quality image generation capability from Stable Diffusion.

It is important to note that the meta design and layer freezing are applied only during Meta ControlNet's training phase. During the adaptation phase, finetuning uses the standard ControlNet training protocol without using layer freezing or meta learning methods.

2.2. Task Selection

Training Tasks: During training phase, we choose HED, Segmentation, and Depth map control as our training tasks. Specifically, we obtain HED map using HED boundary detector proposed by [11]. We obtain Segmentation map using Uniformer [12]. We collect Depth map using Midas [13].

Adaptation Tasks: Follwoing Wang et al. [8], we utilize Canny Edge maps and Normal maps as our adaptation tasks. These tasks, which align the generated image with the control image's edges, are categorized as *edge-based* tasks. Additionally, we introduce two more complex tasks to demonstrate the versatility of our model: Human Pose (line segments to objects) and Human Pose Mapping (objects to line segments), referred to as *non-edge* tasks.

In detail, we collect Canny Edge by using Canny Edge detector [14]. We obtain Normal map by applying Midas [13]. We collect human pose and its reverse human pose mapping by using Openpose [15].

3. Experimental Results

Dataset: We use the generated CLIP-filtered dataset proposed by InstructPix2Pix [16] as our training and validation datasets. The CLIP-filtered dataset contains 313k image-prompt pairs.

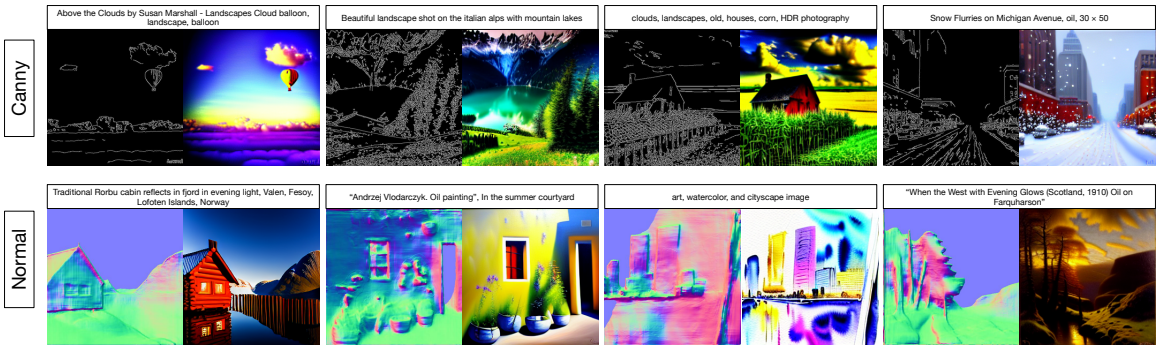

Figure 4: Samples from edge-based tasks (Canny, Normal) in zero-shot adaptation.

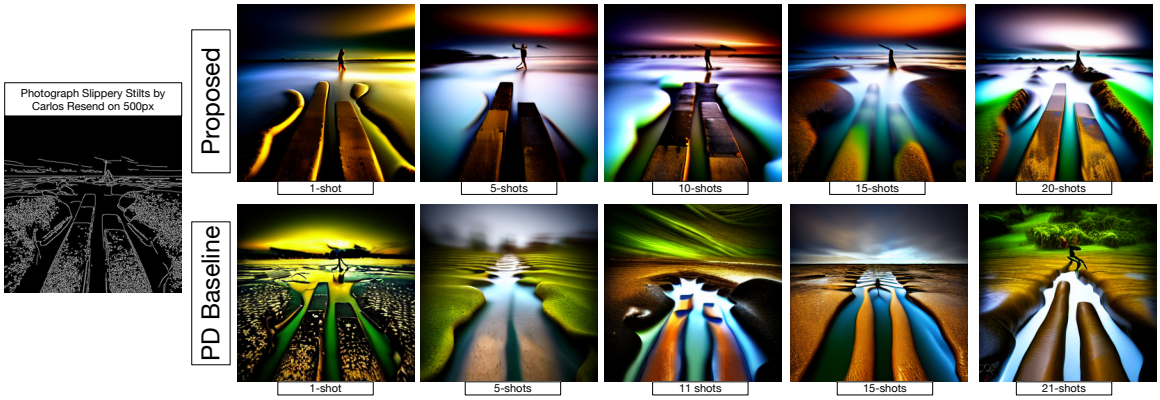

Figure 5: Sample comparison between proposed Meta ControlNet and Prompt Diffusion (PD) baseline for canny task in few-shot finetuning. PD requires example pairs to update and thus is only available in odd number few-shot setting.

Implementations: Meta ControlNet is developed using the ControlNet codebase [7], and utilizes the Stable Diffusion v1.5 checkpoint for finetuning. We fix learning rate to be 1×10^{-4} and batch size to be 256, and accumulate gradients over every 4 batches. Our model is trained on 4 Nvidia A100 GPUs. In our study, both our Meta ControlNet and the baseline of Prompt Diffusion (PD) [8] are evaluated at the 8000-step checkpoint. Note that more finetuning steps enhance image quality.

Regarding the meta design, the meta training images in the inner loop are reused in the meta testing phase in the outer loop for each task in order to optimize memory efficiency. Note that the aforementioned batch size of 256 is the total number of images from all tasks. Namely, we randomly sample 256 images across all tasks, and organize them into batches respectively for each task. This strategy ensures that our algorithm does not require additional image sample during the training phase.

3.1. Fast Control Acquiring in Training

The proposed Meta ControlNet is trained on tasks of HED, segmentation, and depth mapping. Figure 3 displays the validation results after 1000 steps. Clearly, our Meta ControlNet generates the images that closely match the control images with high fidelity. While the vanilla ControlNet requires 5000 steps to exhibit control ability on a single task, our algorithm Meta ControlNet enjoys rapid learning within only 1000 steps. In fact, for most images, control ability of Meta ControlNet occurs within 500 steps or even fewer, with additional training serving to enhance image fidelity.

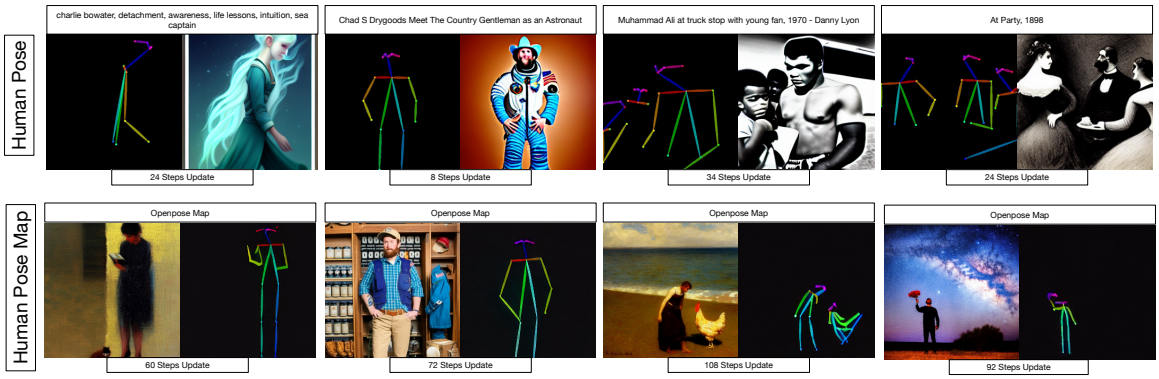

Figure 6: Samples from the validation set for non-edge tasks (Human Pose, Human Pose Map) in a finetuning context. Below each image, the number of updates indicates the first iteration achieving significant control. The validation set is evaluated every two updates.

3.2. Zero-Shot Capability for Edge-Based Tasks

We evaluate Meta ControlNet on edge-based tasks, specifically Canny and Normal tasks. Figure 4 presents the control images alongside the corresponding generated images and text prompts. This adaptation is assessed in a *zero-shot* context, and does not require finetuning or additional data. This is the *first* ControlNet-type method featuring *zero-shot* capability. In contrast, the baseline Prompt Diffusion (PD) method [8] relies on example pairs for learning, rendering it unsuitable for zero-shot settings. The zero-shot results clearly showcase the superior adaptation capability of Meta ControlNet with strong control ability and high fidelity in both tasks.

Further, we compare our Meta ControlNet and the PD baseline with both finetuned in a few-shot context, namely, each method is updated with an equal number of few-shot images. Note that the proposed Meta ControlNet needs only one sample per update step, while PD requires two examples per step. The results in Figure 5 indicate that image quality by Meta ControlNet is enhanced with additional shots. Further, our Meta ControlNet clearly outperforms PD, although the fidelity of the PD gradually improves with more shots. When comparing both methods over the same number of finetuning steps, such as 10 shots for our proposed method and 21 shots for the PD baseline, our approach consistently yields higher quality images. Note that PD requires example pairs to update and thus is only available in odd number few-shot settings.

We highlight that in our experiments, most images generated in zero-shot already exhibit high fidelity, and thus additional few-shot finetuning is not required and might not result in substantial improvements.

3.3. Fast Adaptation for Non-Edge Tasks

We evaluate the generalizability of Meta ControlNet in non-edge tasks, with focus on human pose and its reverse mapping. These non-edge tasks, which are typically challenging in few-shot setting, require a training-like approach for finetuning. To enhance stability, we double the gradient accumulation from 4 to 8 and keep the batch size to be 256.

For Meta ControlNet, we assess validation at every two steps, and record the first instance when the generated image is aligned with the control image, as shown in Figure 6. We observe that for the human pose task, effective control is achieved within 50 steps. The human pose mapping task, which converts human poses into line segments, presents a greater challenge due to the deviation from the high-quality images typically generated by stable diffusion. Nevertheless, Meta ControlNet demonstrates control over most samples within approximately 100 steps. We note that in tasks such as human pose mapping, minor errors can occur, for example, incorrectly depicting a chicken pose

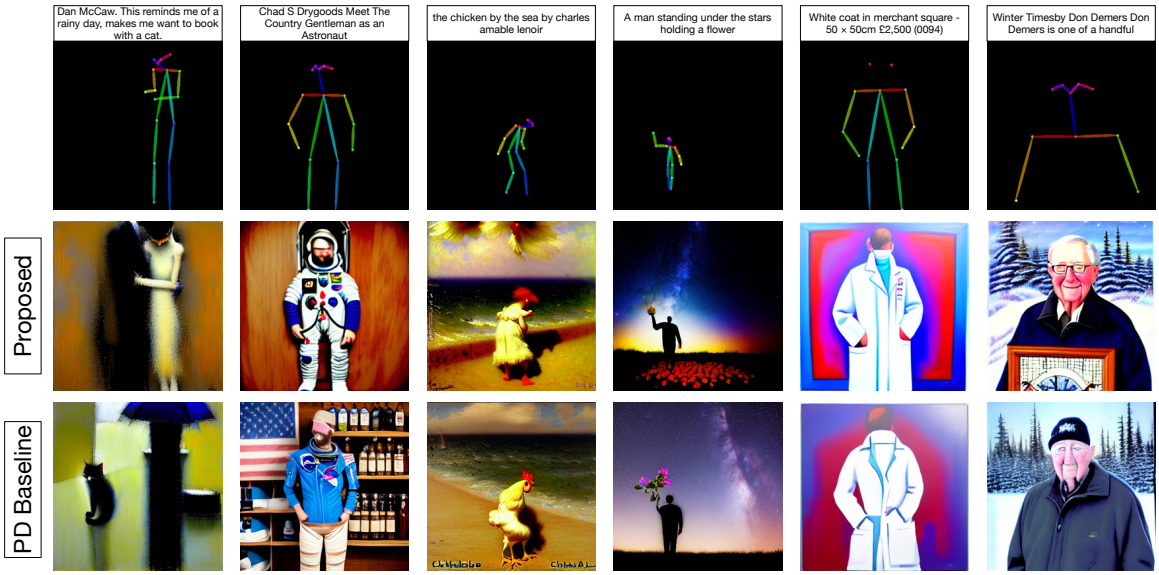

Figure 7: Validation sample comparison between proposed Meta ControlNet and Prompt Diffusion (PD) baseline for Human Pose task after 100 steps of finetuning updates.

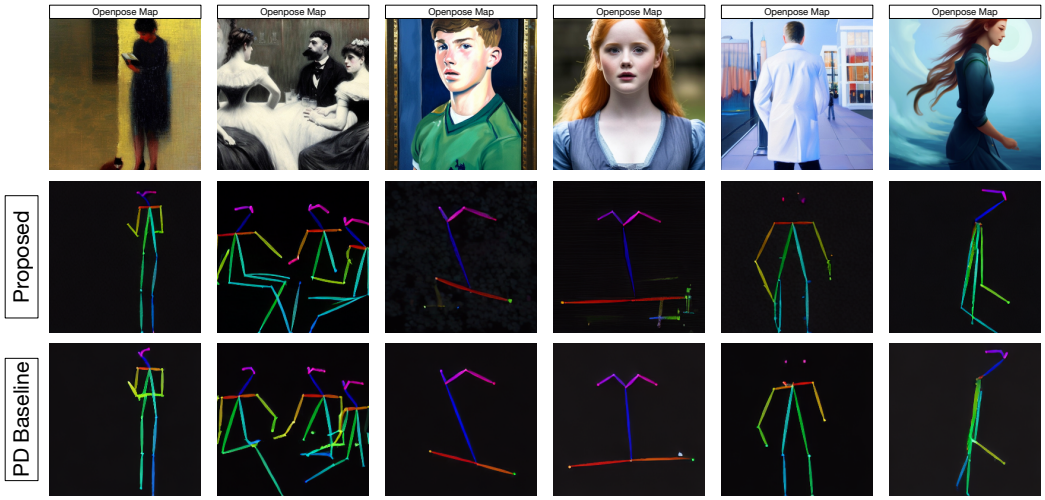

Figure 8: Validation sample comparison between proposed Meta ControlNet and Prompt Diffusion (PD) baseline for Human Pose Mapping task after 200 steps of finetuning updates.

in the third sample pair. This is due to ControlNet's inherent limitation in distinguishing between human and animal, a distinction that requires learning from more samples.

To compare our method with the PD baseline in the human pose task, we evaluate 100-step finetuning in Figure 7. Despite the use of example pairs, PD achieves control but at the cost of reduced fidelity. In contrast, Meta ControlNet maintains both high control and fidelity. For the more challenging human pose mapping task, our method achieves comparable results as the PD baseline with the same 200 steps, but with only half number of images used by PD, demonstrating our better efficiency.

We note that the different convergence speeds to reach control between the human pose task and its mapping counterpart, i.e., 100 versus 200 steps, arise from the inherent characteristics of Stable Diffusion and our choice of training tasks. Both the standard Stable Diffusion and our selected tasks are geared towards generating natural images rather than line segments. Consequently, in the adaptation phase, the model more readily adapts to the human pose task, which involves creating

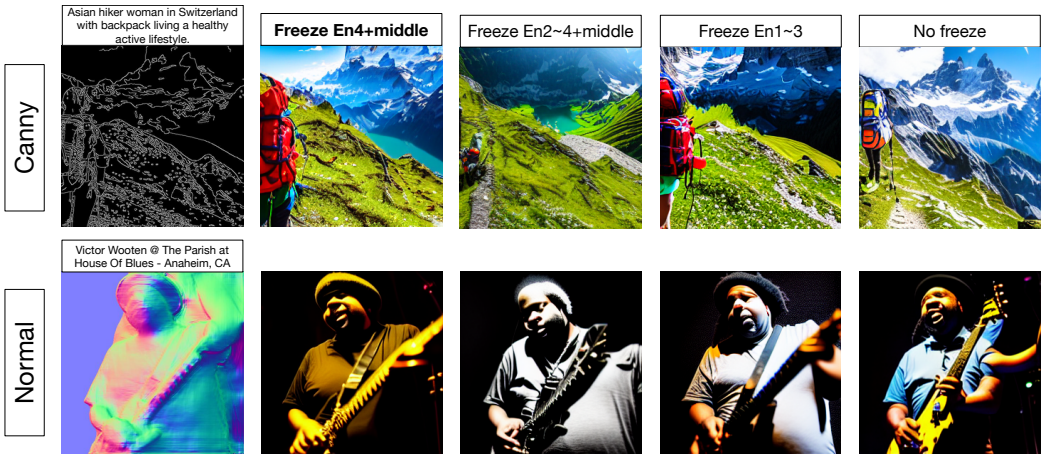

Figure 9: Sample comparison among different freeze methods for edge-based tasks in zero-shot context. En_N refers to the N^{th} Encoder Block. 'Freeze En_4 + Middle' refers to freezing the 4^{th} Encoder Block and the Middle Block in U-Net, which is adapted in Meta ControlNet.

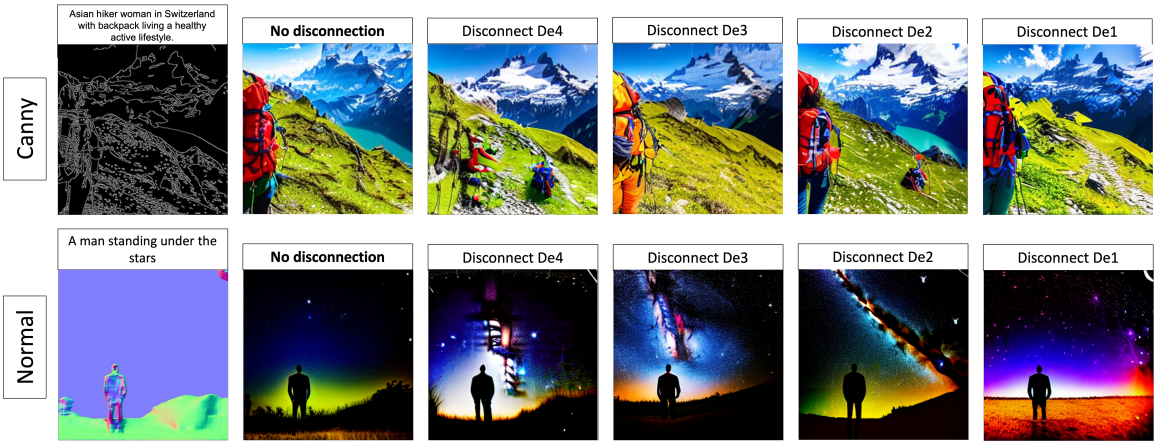

Figure 10: Sample comparison among various connection methods for zero-shot edge-based tasks. De_N refers to the N^{th} Decoder Block.

natural human images, as opposed to learning to generate line segments from human images, a requirement of the mapping task. Nevertheless, our method still achieves better efficiency than PD, namely, comparable results but with only half number of images during training.

3.4. Quantitative results

We quantitatively compare Meta ControlNet with Prompt Diffusion using DreamSim score [17], a visual quality metric derived from Large Vision Models such as DINO [18], CLIP [19], and OpenCLIP [20] that highly aligns with human preferences on visual quality evaluation for both synthetic and real images. We sample the same set of 1130 random images from the test split of InstructPix2Pix [16] for both Prompt Diffusion and Meta ControlNet. Both of the methods also use the same text prompt and the control image for generation. We then measure the average DreamSim score (with different backbone Large Vision Models) over all the generated images and the reference test images for both Meta ControlNet and Prompt Diffusion. We show the Test Average DreamSim Score in Table 1 and Table 2, and Zero-Shot Average DreamSim Score in Table 5 and Table 6. We observe that Meta ControlNet with 1000 training steps achieves significantly better performance than Prompt Diffusion with the same number of training steps over all control tasks from both test

generalization and zero-shot generalization perspectives using any DreamSim variations. The full quantitative comparison results are provided in Appendix A.3.

Test Set Generalization	Average DreamSim Score (ensemble) ↓			Average DreamSim Score (DINO-ViTb16) ↓		
	HED-to-image	Seg-to-image	Depth-to-image	HED-to-image	Seg-to-image	Depth-to-image
Prompt Diffusion (1000 steps)	0.8182	0.8214	0.8354	0.8389	0.8320	0.8498
Meta ControlNet (1000 steps)	0.2905	0.3558	0.3369	0.3225	0.3806	0.3623

Table 1: Test Average DreamSim Scores using Ensemble[1] and Dino-ViTb16 Models as backbone, evaluated under HED, Segmentation, and Depth control tasks.

Test Set Generalization	Average DreamSim Score (CLIP-ViTb32) ↓			Average DreamSim Score (OpenCLIP-ViTb32) ↓		
	HED-to-image	Seg-to-image	Depth-to-image	HED-to-image	Seg-to-image	Depth-to-image
Prompt Diffusion (1000 steps)	0.8338	0.8393	0.8467	0.8296	0.8338	0.8416
Meta ControlNet (1000 steps)	0.2573	0.3182	0.2969	0.2615	0.3217	0.2947

Table 2: Test Average DreamSim Scores using CLIP-ViTb32 and OpenCLIP-ViTb32 Models as backbone, evaluated under HED, Segmentation, and Depth control tasks.

4. Ablation Study

4.1. Layer Freezing

To evaluate the effectiveness of various freezing designs, we conduct an experimental comparison focusing on the canny and normal tasks in a zero-shot setting. The experiment compares our freezing strategy (freezing Encoder Block 4 and the Middle block in U-Net) against other methods: freezing Encoder Blocks 2 to 4 plus the Middle block, freezing Encoder Blocks 1 to 3, and no freezing. The U-Net architecture is available in Appendix. The comprehensive results are detailed in Figure 9, with all methods under identical experimental conditions.

Our results indicate that while all freezing designs facilitate control ability, freezing Encoder Block 4 and the Middle block achieves the most accurate alignment with the control image and the highest image fidelity. This superiority is likely because the first three encoder blocks in U-Net are more task-specific, as they are more directly connected to the control image. In contrast, the latter encoder block and the Middle block contribute significantly to the high-quality image output inherent in stable diffusion, making them ideal candidates for freezing across diverse tasks.

4.2. Decoder Connection

We evaluate our algorithm using different types of connections between the ControlNet decoder and the pre-trained Stable Diffusion (SD) model decoder. Specifically, we use Meta ControlNet with all decoder connected by zero convolution as our baseline. We then disconnect Decoder Blocks 4 to 1 from the pre-trained SD model in separate variants (each Decoder Block n corresponds to Encoder Block n). These variants are trained under the same conditions as the baseline, up to the 8000-step checkpoint, and are then tested on normal and canny edge tasks. Figure 10 depicts the result and suggests that disconnecting Decoder Block 4 significantly reduces image fidelity, and often results in images with repetitive lines or objects with strange shapes. In contrast, disconnecting the other three blocks produces results similar to our baseline. This indicates that Decoder Block 4 plays a critical role in ensuring high image fidelity, which is consistent with our design choice of freezing Encoder Block 4 for this purpose.

5. Conclusion

In our study, we propose a novel Meta ControlNet approach, which adopts the meta learning technique and features novel freezing layer design to learn a generalizable ControlNet initial. This

[1]By default, DreamSim uses an ensemble of CLIP, DINO, and OpenCLIP (all ViT-B/16) to achieve the best human preference alignment on visual quality evaluation.

method exhibits rapid training convergence, and requires only 1000 steps to effectively control generative imaging. Further, such a meta initial exhibits remarkable zero-shot adaptability for edge-based tasks, the first demonstration in this domain. It also excels in more challenging non-edge tasks, and adapts rapidly within 100 steps for the human pose task and 200 steps for the human pose map task. These results not only outperform existing baselines in terms of control ability and efficiency but also represent significant advancement beyond vanilla ControlNet.

Acknowledgments and Disclosure of Funding

The work was supported in part by the U.S. National Science Foundation under the grants ECCS-2113860 and DMS-2134145.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4401–4410, 2019.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[8] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models, 2023.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[10] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[11] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[12] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2023.

[13] RenÃľ Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020.

[14] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

[15] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[16] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[17] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. URL `https://arxiv.org/abs/2306.09344`.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[19] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[20] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[22] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[25] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[28] Denis Zavadski Carsten Rother. Controlnet-xs. *https://vislearn.github.io/ControlNet-XS/*, 2023.

[29] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023.

[30] Ernie Chu, Shuo-Yen Lin, and Jun-Cheng Chen. Video controlnet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models. *arXiv preprint arXiv:2305.19193*, 2023.

[31] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[32] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Yaoyao Liu, Adam Kortylewski, and Alan Yuille. Adding 3d geometry control to diffusion models. *arXiv preprint arXiv:2306.08103*, 2023.

[33] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

[34] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023.

[35] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[36] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

[37] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

[38] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6062–6069, 2020.

[39] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[40] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[41] Baoquan Zhang and Demin Yu. Metadiff: Meta-learning with conditional diffusion for few-shot learning. *arXiv preprint arXiv:2307.16424*, 2023.

[42] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33: 10136–10147, 2020.

[43] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021.

A. Appendix

A.1. Related Work

A.1.1. Diffusion Model and ControlNet

With the recent advancement of score-based generative models [21–24], diffusion models have achieved remarkable performance in text-to-image synthesis [1, 2]. Essentially, the diffusion model learns a time-varying mapping that gradually transforms a random noise into the sample space via a reverse diffusion process [21]. Stable Diffusion (SD) [6], as an important step to achieve high-resolution image generation, utilizes a variational autoencoder to first encode images into a latent space, and then learns a time-conditioned U-Net to perform the denoising process on the latent space based on text prompts.

To allow diffusion models to receive more diverse user-specific guidance for image generation, Composer [25], ControlNet [7], GLIGEN [26] and T2I-Adapter [27] were proposed as general approaches to introduce additional controlling signals. Among these, ControlNet [7] stands out by its superior performance in various downstream tasks ranging from sketch (edge, skeleton) to geometry (depth, normal) guided generation. Technically, ControlNet freezes the original SD model, while finetuning a duplicate of the pre-trained SD to integrate additional control signals via zero-initialized convolution modules. Such an adaptation scheme significantly reduces training costs by re-using the image prior learned in the pre-trained SD. ControlNet-XS [28] further investigates the size and architectural design of ControlNet and proposes a more parameter-efficient architecture. As downstream applications, by leveraging ControlNet, Goel et al. [29] is able to edit the structure and appearance properties for each object in the image, and Chu et al. [30], Wu et al. [31] enforces temporal consistency for video generation, Ma et al. [32], Seo et al. [33] make pre-trained SD aware of 3D knowledge and multi-view geometry. However, for each of these tasks, an independent adapter is required for each condition. The modified version Multi-ControlNet [7] demonstrates the possibility of composing multiple tasks. Uni-ControlNet [34] proposes a unified framework allowing for the simultaneous utilization of different local and global controls. Prompt Diffusion [8] trains an open-domain ControlNet in an in-context learning manner. However, none of these studies are designed for or have been demonstrated to have zero-shot generalization capability for unseen control tasks.

A.1.2. Meta Learning

Meta learning focuses on few-shot learning scenarios, aiming to develop algorithms that leverage a large set of pre-defined tasks to improve performance on unseen instances with only a few or even zero extra training data samples. In this paper, we focus on "learning-to-initialize" approaches such as Model-Agnostic Meta Learning (MAML) [9], which learns an initialization point from which models can fast adapt to new tasks. MAML [9] algorithm involves two layers of training, where at each iteration, the inner loop optimizes the parameter for each task independently starting from the current initialization, whereas the outer loop estimates the gradient with respect to the inner loop optimization path to update the initialization. Since differentiating through an optimization algorithm is often computationally expensive, FO-MAML [9] proposes to simplify the outer-loop gradient computation by directly averaging task-specific gradients evaluated at the outputs of the inner loop. ANIL [10] improves MAML via feature reusing, where the main feature backbone is frozen and only prediction heads are updated by each task in the inner loop. Reptile [35, 36] simplifies the process by aiming for an initialization that minimizes the expected loss across all tasks, similar to joint training.

In computer vision, a popular application of meta learning is few-shot image classification, where a network is adapted to new classes using only a small number of labeled instances per class [9, 37, 38]. Li et al. [39] also adopts MAML-based methods to improve cross-domain generalization. MetaGAN [40] and MetaDiff [41] combines MAML respectively with GAN [5] and

diffusion model [21, 22] to facilitate few-shot image classification. More recently, meta learning has been employed to accelerate implicit neural representation for visual signals [42, 43].

A.2. Detailed Meta ControlNet Architecture

In Figure 11, the detailed Meta ControlNet architecture is illustrated. This architecture shows that during both training and testing phases, the stable diffusion part remain locked, in line with the ControlNet settings. Specifically, for the ControlNet part, SD Encoder Block4 and SD Middle Block are frozen during the meta training phase, while other blocks are subject to finetuning. This approach is chosen because SD Encoder Blocks 1-3 are more closely linked to the control image, necessitating their adaptation to capture task-specific differences. During the meta testing phase, all blocks within the ControlNet part are finetuned to optimize performance.

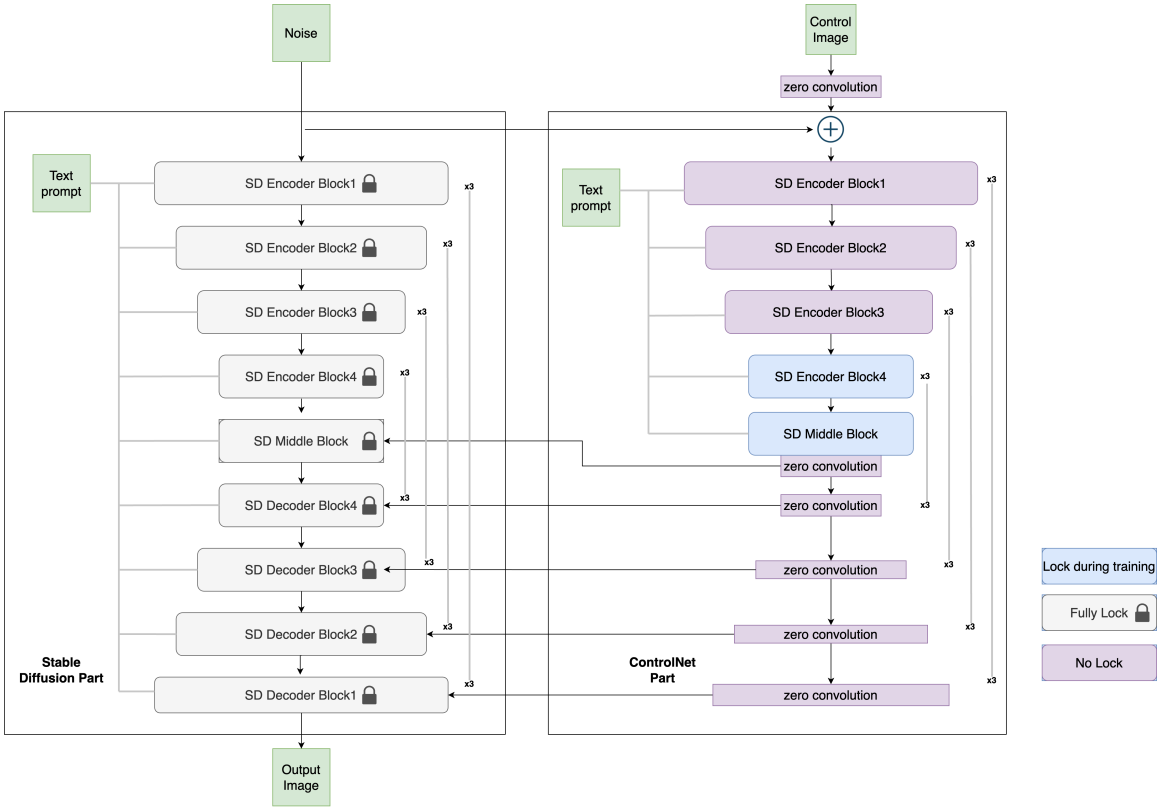

Figure 11: Detailed architecture of Meta ControlNet

A.3. Quantitative results

We quantitatively compare Meta ControlNet with Prompt Diffusion using DreamSim score [17], a visual quality metric derived from Large Vision Models such as DINO [18], CLIP [19], and OpenCLIP [20] that highly aligns with human preferences on visual quality evaluation for both synthetic and real images. We sample the same set of 1130 random images from the test split of InstructPix2Pix [16] for both Prompt Diffusion and Meta ControlNet. Both of the methods also use the same text prompt and the control image for generation. We then measure the average DreamSim score (with different backbone Large Vision Models) over all the generated images and the reference test images for both Meta ControlNet and Prompt Diffusion. We show the Test Average DreamSim Score in Table 1 and Table 2, and Zero-Shot Average DreamSim Score in Table 5 and Table 6. We

observe that Meta ControlNet with 1000 training steps achieves significantly better performance than Prompt Diffusion with the same number of training steps over all control tasks from both test generalization and zero-shot generalization perspectives using any DreamSim variations.

Test Set Generalization	Average DreamSim Score (ensemble) ↓			Average DreamSim Score (DINO-ViTb16) ↓		
	HED-to-image	Seg-to-image	Depth-to-image	HED-to-image	Seg-to-image	Depth-to-image
Prompt Diffusion (1000 steps)	0.8182	0.8214	0.8354	0.8389	0.8320	0.8498
Meta ControlNet (1000 steps)	0.2905	0.3558	0.3369	0.3225	0.3806	0.3623

Table 3: Test Average DreamSim Scores using Ensemble[2] and Dino-ViTb16 Models as backbone, evaluated under HED, Segmentation, and Depth control tasks.

Test Set Generalization	Average DreamSim Score (CLIP-ViTb32) ↓			Average DreamSim Score (OpenCLIP-ViTb32) ↓		
	HED-to-image	Seg-to-image	Depth-to-image	HED-to-image	Seg-to-image	Depth-to-image
Prompt Diffusion (1000 steps)	0.8338	0.8393	0.8467	0.8296	0.8338	0.8416
Meta ControlNet (1000 steps)	0.2573	0.3182	0.2969	0.2615	0.3217	0.2947

Table 4: Test Average DreamSim Scores using CLIP-ViTb32 and OpenCLIP-ViTb32 Models as backbone, evaluated under HED, Segmentation, and Depth control tasks.

Zero Shot Generalization	Average DreamSim Score (ensemble) ↓		Average DreamSim Score (DINO-ViTb16) ↓	
	Canny-to-image	Normal-to-image	Canny-to-image	Normal-to-image
Prompt Diffusion (1000 steps)	0.8179	0.8329	0.8409	0.8580
Meta ControlNet (1000 steps)	0.3358	0.3858	0.3636	0.4081

Table 5: Zero-Shot Average DreamSim Scores using Ensemble and Dino-ViTb16 Models as backbone, evaluated under Canny and Normal map control tasks.

Zero Shot Generalization	Average DreamSim Score (CLIP-ViTb32) ↓		Average DreamSim Score (OpenCLIP-ViTb32) ↓	
	Canny-to-image	Normal-to-image	Canny-to-image	Normal-to-image
Prompt Diffusion (1000 steps)	0.8404	0.8459	0.8261	0.8396
Meta ControlNet (1000 steps)	0.2975	0.3461	0.3017	0.3531

Table 6: Zero-Shot Average DreamSim Scores using CLIP-ViTb32 and OpenCLIP-ViTb32 Models as backbone, evaluated under Canny and Normal map control tasks.

[2]By default, DreamSim uses an ensemble of CLIP, DINO, and OpenCLIP (all ViT-B/16) to achieve the best human preference alignment on visual quality evaluation.