SUPPLEMENTARY MATERIALS
CONTROLAR: CONTROLLABLE IMAGE GENERATION WITH AUTOREGRESSIVE MODELS

Zongming Li[1*], **Tianheng Cheng**[1*], **Shoufa Chen**[2], **Peize Sun**[2], **Haocheng Shen**[3], **Longjin Ran**[3], **Xiaoxin Chen**[3], **Wenyu Liu**[1] **& Xinggang Wang**[1†]
[1] School of EIC, Huazhong University of Science and Technology
[2] Department of Computer Science, The University of Hong Kong
[3] vivo AI Lab

A APPENDIX

A.1 IMPLEMENTATION DETAILS

Dataset details. The quantity of images from all datasets utilized in our experiment is detailed in Tab. 1. We utilize the ImageNet-1K (Deng et al., 2009) as the training dataset for class-to-image controllable generation, encompassing a total of 1,000 classes. The canny edge detector (Canny, 1986) is employed to acquire the canny edge map, and the depth map is obtained using Midas (Ranftl et al., 2020). In the context of text-to-image controllable generation, ADE20K (Zhou et al., 2017) and COCOStuff (Caesar et al., 2018) are harnessed for training the segmentation control task, while MultiGen-20M is utilized for training the edge map and depth control generation.

Table 1: **Details of different dataset.**

	ImageNet-1K	ADE20K	COCOStuff	MultiGen-20M
Training Samples	1281188	20210	118287	2810616
Evaluation Samples	50000	2000	5000	5000

Evaluation details. To assess the conditional consistency of the generated images, we have devised various metrics tailored to each specific task. In the context of segmentation control generation, we employ a segmentation model to evaluate the mean Intersection over Union (mIoU) of the generated images. Specifically, we reference ControlNet++ to examine the results of the validation set generation on ADE20K using Mask2Former (Cheng et al., 2022), and on COCOStuff using DeepLabv3 (Chen, 2017). For canny edge control generation, we utilize the canny edge detector with thresholds of (100, 200) to derive the canny edge of the results, and subsequently calculate the F1-Score in relation to the input control. In the case of hed and lineart edge, we follow the approach outlined in ControlNet to obtain control images and compute the Structural Similarity Index (SSIM). Regarding depth map control generation, we calculate the Root Mean Square Error (RMSE).

Table 2: **Training details of different tasks.**

	Seg.		Canny	Hed	Lineart	Depth
	ADE20K	COCOStuff	MultiGen-20M			
Batch size	96	96	96	88	88	96
GPU hours	55	80	340	160	110	370

Training details. We use 8 Nvidia A100 80G GPUs to complete text-to-image controllable generation experiments based on LlamaGen-XL (Sun et al., 2024). The batch size settings and GPU hours

*Equal contributions; †Corresponding author: xgwang@hust.edu.cn.

during training can be found in Tab. 2. We use the edge extraction model to obtain the hed edge and lineart edge of the image during the training process, which takes up some memory, so the batch size is slightly smaller than the other tasks. It should be noted that since the ADE20K dataset has less training data, we first merge the ADE20K and COCOStuff datasets together to train the model, which requires roughly 50 GPU hours. Because the segmentation map labelling is inconsistent between the two datasets, we fine-tuned 2k iterations on ADE20K and 20k iterations on COCOStuff, respectively. The additional 2k iteration on ADE20K results in a mIoU improvement of 1.15.

A.2 More experimental explorations

Comparison with recent work. We have added some quantitative comparative results with recent work including OmniGen (Xiao et al., 2024) and Lumina-mGPT (Liu et al., 2024), as shown in the Tab. 3. The results for segmentation task are measured on the validation set of ADE20K (Zhou et al., 2017), and the results for canny, hed and depth are measured on the validation set of MultiGen-20M (Qin et al., 2023). OmniGen uses iterative denoising diffusion for image generation, while lumina-mGPT uses autoregressive prediction. Although Lumina-mGPT has a much larger number of parameters than our ControlAR, it does not perform particularly well on the controllable generation task. Our ControlAR provides a good solution for autoregressive controllable image generation and our method does not require any adjustments to the structure of the generative network or modifications to the length of the sequences, which means that we can easily migrate our ControlAR to other autoregressive image generation models, such as Lumina-mGPT.

Table 3: **Quantitative comparison with recent works.**

Method	Param.	Seg.(mIoU↑)	Canny(F1-Score↑)	Hed(SSIM↑)	Depth(RMSE↓)
OmniGen	3.8B	44.23	35.54	82.37	28.54
Lumina-mGPT	7B	25.02	29.99	78.21	55.25
Ours	0.8B	39.95	37.08	85.63	29.01

Adjustable control strength. Given the diversity of image structures, we sometimes do not want the spatial structure of the generated image to be identical to the input control. To achieve this, it is only necessary to skip the operation of fusing the control condition token with the image token with a probability of 50% when training ControlAR. Such an approach ensures ControlAR's generative capability in the absence of control image inputs. At the same time, multiplying the control condition token by a control strength factor α during inference changes the degree of control of the generated result. When α is 1, ControlAR will generate an image exclusively based on the control condition, while when α is 0, the generated results will be related only to the text prompt. Fig. 1 shows the visualizations using edges as the control image and adjusting the control strength.

Arbitrary-Resolution Generation Without Condition Image. We conduct a more in-depth exploratory study on resolution control in the absence of specific condition image. We can generate a grayscale map of the corresponding resolution according to the desired height and width, this grayscale map consists of a number of 16 × 16 small squares, and the grayscale value of each row decreases from left to right, the left most 255, the right most 0. This grayscale image is the condition image that determines the resolution. Thanks to the strong positional dependence of the control decoding strategy between the image token and the control condition token, the model only needs to generate a sequence as long as the control condition sequence. And since the grayscale value of each row is decreasing from left to right, the model can easily know when it is necessary to switch to the next row. We have verified the feasibility of this approach on a small experimental scale. We show some visualization results in Fig. 2. Using resolution-aware prompts to control the resolution as in Lumina-mGPT requires the constant generation of `<end-of-line>` tokens during the prediction of the image and the eventual prediction of `<end-of-image>` token. This

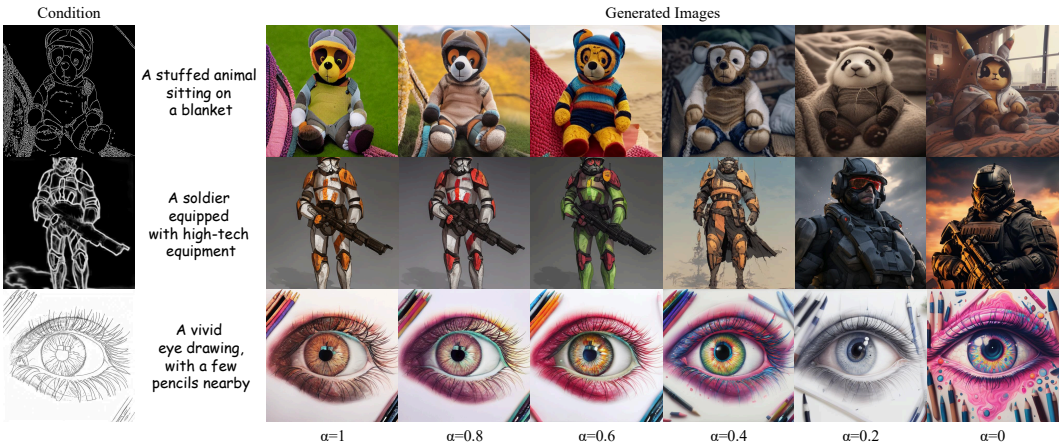

Figure 1: **Visualization with different control strength factor α.**

approach requires the model to make its own decisions about where to make line breaks and where to end generation, but our ControlAR is directly telling the model where to make line breaks and end generation. We only need to fine-tune the weights based on LlamaGen-XL (512×512) on about 1M text-image paired data for 30k steps to achieve a good arbitrary resolution generation capability without specific control image. This proves that our ControlAR can be a very effective strategy for controlling resolution.

A.3 Discussion

Limitation. We have shown in our experiments that updating the parameters of the generative model can achieve better results than freezing it completely. However, this approach is still not as convenient as ControlNet in terms of model portability. In addition, our method does not currently support scenarios where multiple control images are input simultaneously. Processing multiple control images simultaneously using a control encoder with a small number of parameters can be challenging.

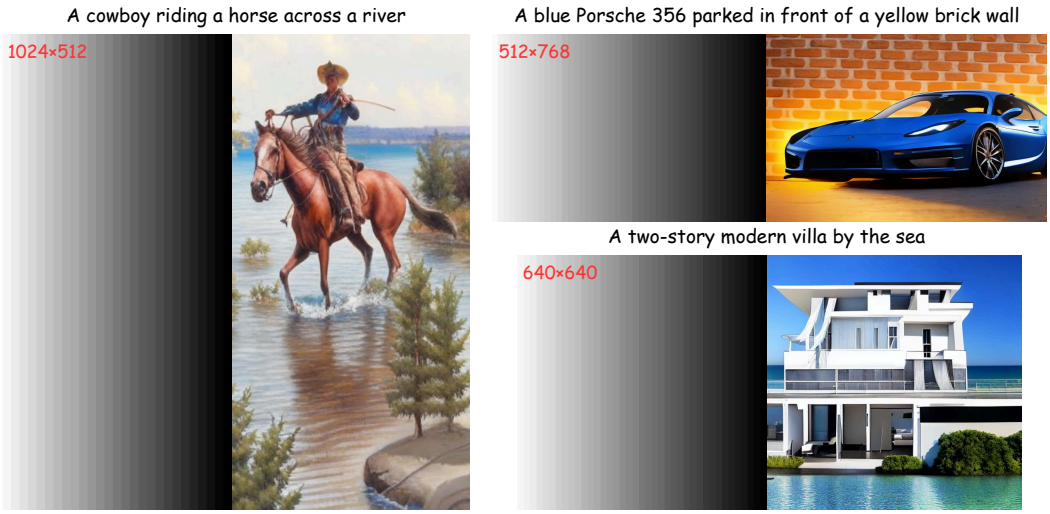

Figure 2: **Arbitrary-Resolution generation without condition image.** The grayscale map on the left is the condition image generated according to the desired resolution.

Failure Cases. ControlAR performs well on the conditional consistency of controllable generation of spatial structures. But because of this, the generated images are sometimes less controlled by the text prompt, especially when the textual prompt conflicts with the spatial structure of the control image. We use depth-to-image and canny-to-image as examples in Fig. 3. When there is a large difference between the text prompt and the original image, it might fail to generate images according to the text prompt. In ControlAR, we can use the control factor to adjust the strength of spatial control, thereby aligning the generated results with the text and mitigating this conflict. However, the conflict between text prompts and spatial controls is a common issue in current control-to-image generation models, including ControlNet (Zhang et al., 2023) and ControlNet++ (Li et al., 2024). As shown in Fig. 3, neither ControlNet nor ControlNet++ can generate images that strictly follow the text prompts. Moreover, ControlNet++ introduces additional supervision to facilitate alignment between the generated image and spatial controls, which weakens the influence of the text prompt as shown in the case of canny-to-image. This phenomenon reflects that there may be some confrontation between the structural freedom of the generated image and the conditional consistency. The examples in Fig. 3 reflect the possible contradiction between structure diversity and conditional consistency. We acknowledge that structure diversity is a meaningful and challenging problem for controllable image generation. We extend ControlAR by introducing a control α to dynamically adjust the strength of control. This allows the model to balance structural consistency and diversity, enabling the generated images to align with the input geometric controls while also introducing variations to produce richer and more diverse structures. Although this is an exploratory attempt, we believe that ControlAR has the potential to achieve this balance. Specifically speaking, in order to improve the diversity of generated images, we believe that we need to explore suitable training strategies to achieve the effect of being able to adjust the intensity of control during the inference phase, and to resolve the possible contradiction between text alignment and conditional consistency, which are important directions in future research.

Figure 3: **Failure cases of current ControlAR.** When the text prompt conflicts with the control image, the generated result tends to ignore the text prompt. Adjusting the control strength factor α can alleviate this problem.

Future work. We will use more data to try more kinds of conditional control generation, such as human pose and bounding box. At the same time, in order to improve the migratability of the model we will consider focusing the parameter update on the control encoder and keep the parameters of the generated model itself unchanged. In addition to this, how to use one control encoder to process different control image inputs simultaneously is also a direction worth exploring.

A.4 More Visualizations

More visualization results under different conditions of control are shown in Fig. 4 5 6 7 8. We alse show some visualization comparison of ControlAR and MR-ControlAR at different resolution in Fig. 9 and Fig. 10.

Figure 4: **Segmentation mask control generation visualization.**

Figure 5: **Canny edge control generation visualization.**

Figure 6: **Hed edge control generation visualization.**

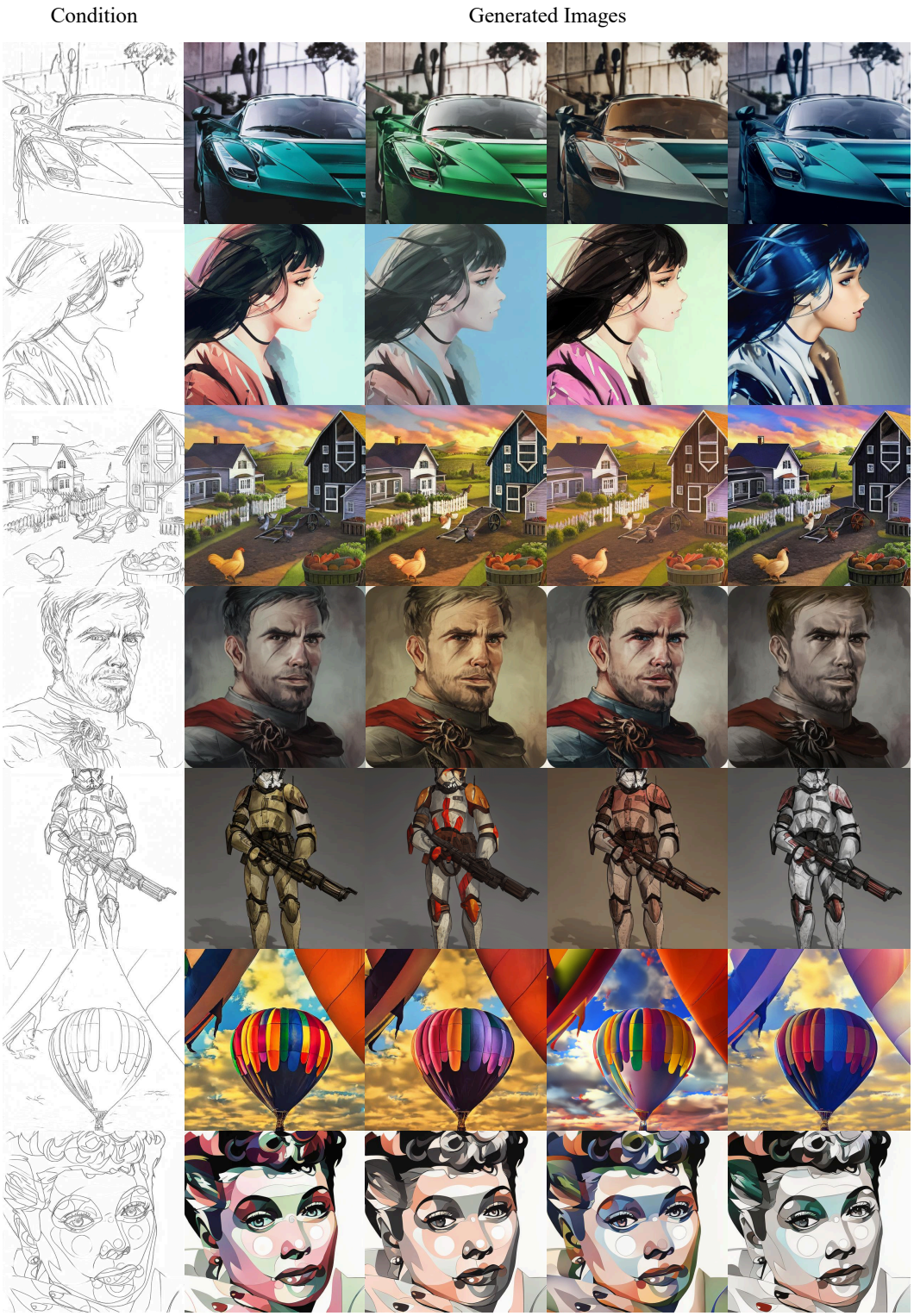

Figure 7: **Lineart edge control generation visualization.**

Figure 8: **Depth map control generation visualization.**

Condition | MR-ControlAR (SSIM：86.98) | ControlAR (SSIM：79.21)

Figure 9: **visualization comparison of MR-ControlAR and ControlAR at the resolution of** 1024×512**.**

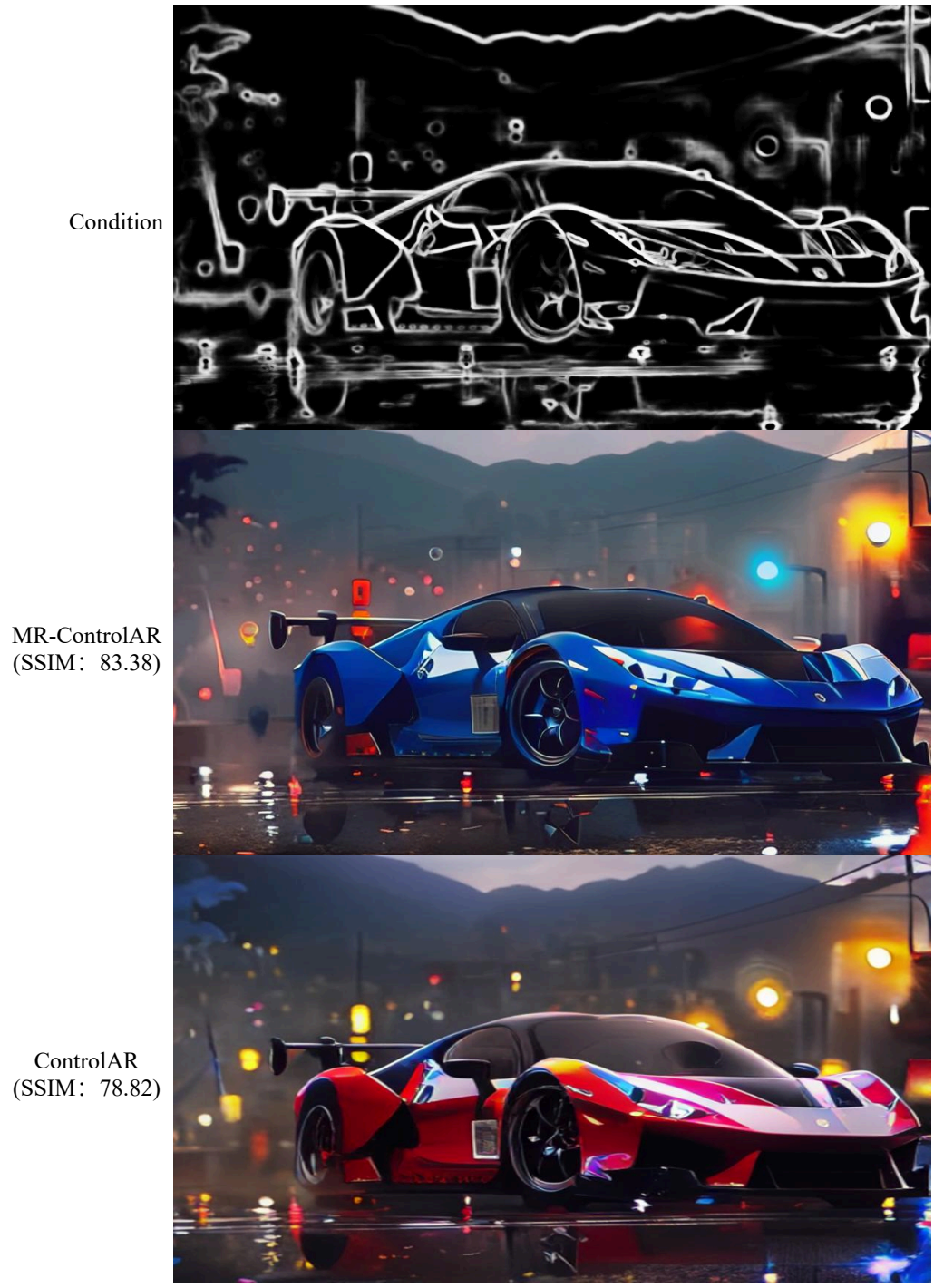

Figure 10: **visualization comparison of MR-ControlAR and ControlAR at the resolution of** 576×1024**.**

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 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, 2020.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

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

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.