INTRINSIC LORA: A Generalist Approach for Discovering Knowledge in Generative Models

Xiaodan Du[1] Nicholas Kolkin[2] Greg Shakhnarovich[1] Anand Bhattad[1]
[1]Toyota Technological Institute at Chicago [2]Adobe

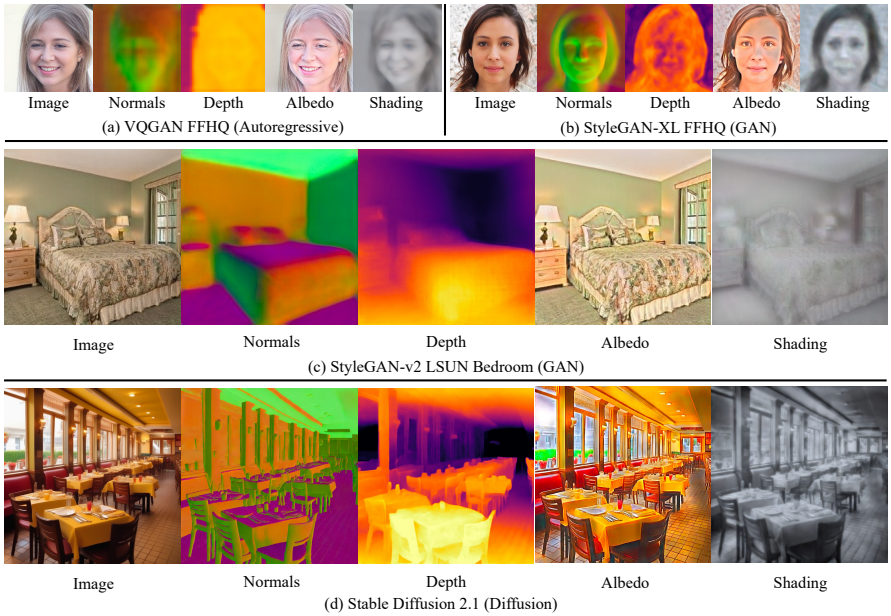


Figure 1. INTRINSIC LORA (I-LORA) is a general approach for extracting visual knowledge from generative models of many types. Our method applies targeted, lightweight fine-tuning to modulate key feature maps, using low-rank adaptation (LoRA) on attention layers in VQGAN (a) and Stable Diffusion (d), and affine layers in StyleGAN (b and c). This process helps us discover fundamental scene intrinsics – normals, depth, albedo, and shading – directly from the models' learned representations, avoiding the need for additional task-specific design of decoding heads or layers.

Abstract

Generative models have been shown to be capable of creating images that closely mimic real scenes, suggesting they inherently encode scene representations. We introduce INTRINSIC LORA *(*I-LORA*), a general approach that uses Low-Rank Adaptation (LoRA) to discover scene intrinsics such as normals, depth, albedo, and shading from a wide array of generative models.* I-LORA *is lightweight, adding minimally to the model's parameters and requiring very small datasets for this knowledge discovery. Our approach, applicable to Diffusion models, GANs, and Autoregressive models alike, generates intrinsics using the same output head as the original images.*

1. Introduction

Generative models can produce high-quality images almost indistinguishable from real-world photographs. They seem to demonstrate a profound understanding of the world, capturing nuances of realistic object placement, appearance, and lighting conditions. Yet, it remains an open question how these models encode such detailed knowledge, and whether representations of scene intrinsics exist in these models and can be extracted explicitly.

Table 1. Summary of scene intrinsics found across different generative models without changing generator head. ✓: Intrinsics can be extracted with high quality. ∼: Intrinsics cannot be extracted with high quality. ×: Intrinsics cannot be extracted.

Model	Pretrain Type	Domain	Normal	Depth	Albedo	Shading
VQGAN [14]	Autoregressive	FFHQ	∼	∼	✓	✓
SG-v2 [27]	GAN	FFHQ	✓	∼	✓	✓
SG-v2 [54]	GAN	LSUN Bed	✓	✓	✓	✓
SG-XL [44]	GAN	FFHQ	✓	∼	✓	✓
SG-XL [44]	GAN	ImageNet	×	×	×	×
SD-UNet (single-step) [42]	Diffusion	Open	✓	✓	✓	✓
SD (multi-step) [42]	Diffusion	Open	✓	✓	✓	✓

Our Contribution. We conduct our inquiry across a spectrum spanning diffusion, GANs, and autoregressive models – to understand whether they encode fundamental scene in-

trinsics of normals, depth, albedo, and shading [3]. Our method, INTRINSIC LORA (I-LORA), a Low-Rank Adaptation (LoRA) technique, efficiently extracts these intrinsics across different model types with minimal computational overhead and data requirements. Detailed results and a summary are presented in Tab. 1 and elaborated further in Sec. 4. Our experiments suggest that the intrinsic knowledge within generative models is not accidental but a byproduct of large-scale learning to mimic image data. In summary, our work broadens the understanding of visual knowledge within generative image models and our contributions are:

- **Wide Applicability:** We validate I-LORA 's capability to extract scene intrinsics (normals, depth, albedo, and shading) across a broad spectrum of generative models, highlighting its adaptability to diverse architectures.
- **Efficient and Lean Approach** to knowledge extraction: I-LORA is highly efficient, requiring a little increase in parameters (less than 0.17% for Stable Diffusion) and minimal training data, as few as 250 images.
- **Insights from Learned Priors:** Through control experiments, we illustrate the critical role of learned priors, suggesting the quality of intrinsics extracted is correlated to the visual quality of the generative model.
- **Competitive Quality of Intrinsics:** Our method, supervised with hundreds to thousands of labeled images, generates intrinsics on par with or even better than those produced by the leading supervised techniques requiring millions of labeled images.

2. Related Work

Generative Models: Generative Adversarial Networks (GANs) [17] have been widely used for generating realistic images. Variants like StyleGAN [25], StyleGAN2 [27] and GigaGAN [23] have pushed the boundaries in terms of image quality and control.

Diffusion models, such as Denoising Score Matching [49] and Noise-Contrastive Estimation [18], have been used for generative tasks and are perhaps the most popular at the moment [20, 28, 42].

Autoregressive models like PixelRNN [47] and PixelCNN [46] generate images pixel-by-pixel, offering fine-grained control but at the cost of computational efficiency. More recently, VQ-VAE-2 [41] and VQGAN [14] have combined autoregressive models with vector quantization to achieve high-quality image synthesis.

Scene Intrinsics Extraction: Barrow and Tenenbaum [3] highlighted several fundamental scene intrinsics including depth, albedo, shading, and surface normals. A large body of work has focused on extracting some related properties, like depth and normals from images [4, 12, 13, 24, 32, 40] using labeled annotated data. Labeled annotations of albedo and shading are hard to find and as the recent review in [15] shows, methods involving little or no learning have remained competitive until fairly recently. However, these methods often rely on supervised learning and do not explore the capabilities of generative models in this context.

Many recent studies have used generative models [1, 2, 22, 29, 30, 37, 43, 51, 57, 58] as pre-trained feature extractors or scene prior learners. They use generated images to enhance downstream discriminative models, fine-tune the original generative model for a new task, learn new layers or decoders to produce desired scene intrinsics.

Knowledge in Generative Models: Several studies have explored the extent of StyleGAN's knowledge, particularly in the context of 3D information about faces [38, 56]. Yang et al. [52] show GANs encode hierarchical semantic information across different layers. Further research has demonstrated that manipulating offsets in StyleGAN can lead to effective relighting of images [5] and extraction of scene intrinsics [7]. Chen et al.[9] found internal activations of the LDM encode linear representations of both 3D depth data and a salient-object / background distinction. Recently, [19, 34, 45] found correspondence emerges in image diffusion models without any explicit supervision.

LoRA (Low-Rank Adaptation). LoRA [21] introduces trainable low-rank decomposed matrices into specific layers of the model architecture. These matrices are the only components updated during task-specific optimization. This results in a significant reduction in the number of trainable parameters, ensuring only slight modifications to the model, and preserving its core functionality and accessibility.

3. INTRINSIC LORA

A generative model G maps noise/conditioning information z to an RGB image $G(z) \in \mathbb{R}^{H\times W\times 3}$. We seek to augment G with a small set of parameters θ that allow us to produce, using the same architecture as G, an image-like map with up to three channels, representing scene intrinsics.

I-LORA's Learning Framework. Our method, I-LORA, learns to extract intrinsic properties of an image (such as depth) using a small number of labeled examples (image/depth map pairs) as supervision. In cases where we do not have access to the actual intrinsic properties, we use models trained on large datasets to generate estimated intrinsics as pseudo-ground truth, used as training targets for G_θ.

To optimize θ of G_θ using a pseudo-ground truth predictor Φ (e.g., a network trained to predict depth from an image), we minimize the objective:

$$\min_\theta \mathbb{E}_z[d(G_\theta(z), \Phi(G(z)))], \tag{1}$$

where d is the distance metric.

Diffusion models require special treatment since they are effectively image-to-image and not noise-to-image. During inference, diffusion models repeatedly receive a noisy image as input. Thus instead of conditioning noise z we feed an

image x (generated or real) to a diffusion model G. In this case, given a real image x, our objective function becomes $\min_\theta \mathbb{E}_x[d(G_\theta(x), \Phi(x))]$.

For surface normals Φ is Omnidatav2-Normal [12, 24]. For depth we use ZoeDepth [4] as the predictor Φ. For Albedo and Shading Φ is Paradigms [6, 15]. For SG2, SGXL and VQGAN, d in Eq.1 is

$$d(x, y) = 1 - cos(x, y) + \|x - y\|_1 \tag{2}$$

for normal and MSE for other intrinsics. For latent diffusion based methods, there isn't a clear physical meaning to the relative angle of latent vectors in encoded normal maps, so we use the standard objective of MSE for all intrinsics.

We use LoRA to recover image intrinsics from generative models. LoRA introduces a low-rank weight matrix W^*, which has a lower rank than the original weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$. This is achieved by factorizing W^* into two smaller matrices $W_u^* \in \mathbb{R}^{d_1 \times d^*}$ and $W_l^* \in \mathbb{R}^{d^* \times d_2}$, where d^* is chosen such that $d^* \ll \min(d_1, d_2)$.

The output o for an input activation a is then given by:

$$o = Wa + W^*a = Wa + W_u^* W_l^* a. \tag{3}$$

Applying I-LoRA. For **GANs**, I-LoRA modules are integrated with the affine layers that map from w-space to s-space [50]. In the case of **VQGAN, an autoregressive model**, I-LoRA is applied to the convolutional attention layers within the decoder. For **diffusion models**, I-LoRA adaptors are learned atop cross-attention and self-attention layers. The UNet is utilized as a dense predictor, transforming an RGB input into intrinsics in one step. This approach, favoring simplicity and effectiveness, delivers superior quantitative results. Depending on the intrinsics of interest, the textual input varies among "surface normal", "depth", "albedo", or "shading".

4. Experiments

In this section, we outline I-LoRA's contributions, demonstrating its general applicability across generative models (Sec. 4.1). Control experiments provide evidence of I-LoRA's effectiveness (Sec. 4.2). Note: our analysis in Sec. 4.2 primarily utilizes a single-step I-LoRA model for intrinsic image extraction. In Sec. 5, we discuss the challenge of naively applying I-LoRA to a multi-step Stable Diffusion model. We propose a simple modification to the architecture by adding an extra layer (that is not learned) for improved intrinsic image extraction. We refer to this model as **Augmented I-LoRA** (I-LoRA$_{\text{AUG}}$).

4.1. I-LoRA is General and Universally Applicable

We evaluate I-LoRA across diverse generative models, including StyleGAN-v2 [55], StyleGAN-XL [44], and VQGAN [14], trained on datasets like FFHQ [27], LSUN Bedrooms [53], and ImageNet [11]. I-LoRA adaptors are tailored to each model and dataset to extract intrinsics: surface normals, depth, albedo, and shading, demonstrating broad applicability and robustness in both qualitative assessments (Fig. 1, 2, 4) and quantitative (Tab. 2 on generated images, Tab. 3 on real images). In all experiments – covering both generated and real images – we use pseudo-ground truth from off-the-shelf models as a supervisory signal. We use I-LoRA with Rank 8 as default for all generative models.

We find I-LoRA can unearth intrinsic knowledge across almost all models tested, the notable exception is StyleGAN-XL trained on ImageNet. Where it yields qualitatively poor results, which we attribute to the model's limited ability to generate realistic images (Fig. 3). This suggests the quality of intrinsic extraction is correlated with the generative model's fidelity (see Sec. 4.2).

In evaluations of generated images, our method is benchmarked against pseudo-ground truths derived from existing models, compensating for the lack of true ground truths.

Diffusion models excel as powerful image generators, thanks to their architecture as image-to-image translators. This feature simplifies their application to real images. Tak-

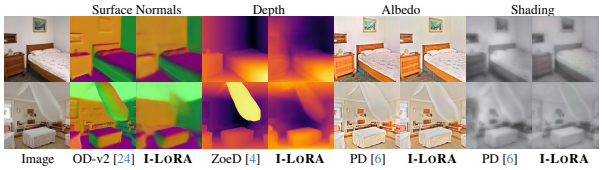


Figure 2. Scene intrinsic properties extracted from StyleGAN-v2 trained on LSUN bedroom images using I-LoRA.

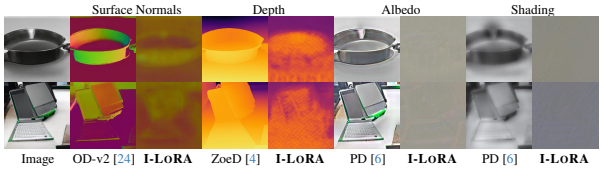


Figure 3. StyleGAN-XL trained on ImageNet. Top: pan, bottom: laptop, with the corresponding scene intrinsics (pseudo ground truth and extracted) alongside. The surface normals and depth maps, while capturing the basic shape and volume, lack precise detail and exhibit artifacts. Albedo and Shading extractions fail. These difficulties are correlated with the overall worse realism and consistency of the generated images.

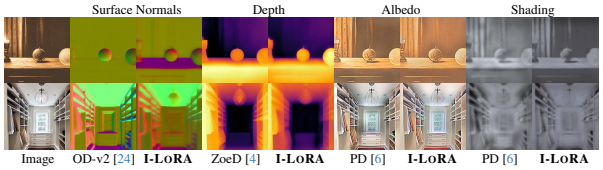


Figure 4. Scene intrinsics from I-LoRA applied to randomly generated images. I-LoRA accurately predicts the table's normal in the first row when compared to [24]. The comparison highlights I-LoRA's ability to closely align with, and sometimes surpass, these supervised SOTA monocular predictors.

Table 2. Quantiative analysis of scene intrinsics extraction performance by I-LoRA on generated images. We compare with pseudo GT from Omnidatav2-normal, ZoeDepth and Paradigms.

Model	Pre-training Type	Domain	LoRA Param.	Surface Normal			Depth		Albedo	Shading
				Mean Error°↓	Median Error°↓	L1 Error $_{\times 100}$ ↓	RMS ↓	$\delta < 1.25_{\times 100}$ ↑	RMS ↓	RMS ↓
VQGAN	Autoregressive	FFHQ	0.18%	19.97	20.97	16.33	0.1819	62.33	0.0345	0.0106
StyleGAN-v2	GAN	FFHQ	0.57%	16.93	19.60	13.87	0.1530	90.74	0.0283	0.0110
StyleGAN-XL	GAN	FFHQ	0.29%	15.28	18.07	12.63	0.1337	93.87	0.0287	0.0125
StyleGAN-v2	GAN	LSUN Bedroom	0.57%	13.94	24.76	11.49	0.0897	66.88	0.0270	0.0074
StyleGAN-XL	GAN	ImageNet	0.29%	24.09	25.52	19.44	0.2175	38.38	0.1065	0.0119
I-LoRA$_{\text{AUG}}$ (multi step)	Diffusion	Open	0.17%	21.41	28.57	17.39	0.2042	41.21	0.0881	0.0099
I-LoRA (single step)	Diffusion	Open	0.17%	16.63	23.64	13.69	0.1179	52.59	0.0487	0.0118

Table 3. Quantitative analysis of scene intrinsic extraction performance across different models on real images.

Model	Pre-training Type	LoRA Param	Surface Normal			Depth	
			Mean Error°↓	Median Error°↓	L1 Error $_{\times 100}$ ↓	RMS ↓	$\delta < 1.25_{\times 100}$ ↑
Omnidata-v2 [24]/ZoeDepth [4]	Supervised	-	**18.90**	13.36	**15.21**	0.2693	**47.56**
I-LoRA$_{\text{AUG}}$ (multi step)	Diffusion	0.17%	23.74	19.08	19.31	0.2651	43.19
I-LoRA (single step)	Diffusion	0.17%	20.31	**12.54**	16.53	**0.2046**	44.90

Figure 5. We find a correlation between generative model quality and scene intrinsic extraction accuracy.

ing advantage of this, we apply I-LoRA to directly extract intrinsic images from Stable Diffusion's UNet in a single step. This method bypasses the iterative reverse denoising process. The model receives a real image as input and outputs the corresponding image intrinsics through I-LoRA. Such direct application allows for evaluation against actual ground truth. This provides a definitive benchmark for assessing I-LoRA's effectiveness (Tab. 3 and Fig. 5). We evaluate on DIODE dataset [48] containing a diverse range of complex indoor and outdoor scenes.

In Tab. 3, we find that I-LoRA not only matches but, in several metrics (surface normals median error, depth RMSE), surpasses the performance of Omnidata and ZoeDepth – the source of its training signal – while using significantly less data, parameters, and training time.

4.2. Control Experiments and Correlation with Generative Quality

To assess if our I-LoRA leverages pre-trained generative capabilities or primarily depends on LoRA layers, we performed a control experiment using a randomly initialized SD UNet, following the same training protocol of our I-LoRA model. The poor results from this model, presented in Fig. 5, corroborate that the learned features developed during generative pre-training are crucial for intrinsic extraction, rather than I-LoRA layers alone.

Furthermore, analyzing multiple Stable Diffusion versions (v1-1, v1-2 and v1-5) under the same training protocol reveals that enhancements in image generation quality correlate positively with intrinsic extraction capabilities. This

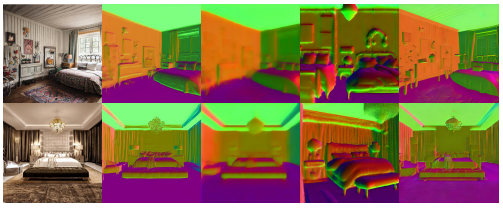


Figure 6. I-LoRA yields satisfactory results, but multiple diffusion steps lead to misalignment in extracted intrinsics, see the SDv1-5 column. The last column, I-LoRA$_{\text{AUG}}$, demonstrates successfully correcting the misalignment using our image conditioning approach, resulting in well-aligned and detailed intrinsic extractions

assertion is further reinforced by observing a correlation between lower FID scores (9.6 for VQGAN [14], 3.62 for StyleGAN-v2 [26] and 2.19 for StyleGAN-XL [44]) and improved intrinsic predictions in our FFHQ dataset experiments, illustrated in Tab. 2 (first three rows).

5. I-LoRA$_{\text{AUG}}$: Augmented I-LoRA

Can we enhance the quality of the intrinsics by leveraging the multi-step diffusion inference? While multi-step diffusion improves sharpness, we find it introduces two challenges: 1. intrinsics misaligned with input, and 2. shift in the distribution of outputs relative to the ground truth (visually manifesting as a color shift) (see Fig. 6).

To address the first challenge, we augment the noise input to the UNet with the input image's latent encoding, as in InstructPix2Pix [8] (IP2P). The second challenge is a known artifact attributed to Stable Diffusion's difficulty generating images that are not with medium brightness [10, 31]. Following [31], we replace SDv1-5 with SDv2-1 while maintaining our previously described learning protocol. We name this multi-step augmented SDv2-1 model I-LoRA$_{\text{AUG}}$. I-LoRA$_{\text{AUG}}$ solves the misalignment issue and reduces the color shift significantly (Fig. 6), resulting in the generation of high-quality, sharp scene intrinsics with improved quantitative accuracy. However, quantitatively, the results still fall short of our single-step I-LoRA result. In the future, we hope this problem will be solved by improved sampling techniques and the next generation of generative models.

6. Conclusion

In conclusion, we find consistent evidence that generative models implicitly learn scene intrinsics, allowing tiny LoRA adaptors to extract this information with minimal fine-tuning on small labeled data. More powerful generative models produce more accurate intrinsics, strengthening our hypothesis that learning this information is a natural byproduct of learning to generate images well. Finally, we discovered scene intrinsics exist across different generative models, resonating with Barrow & Tenenbaum's hypothesis of fundamental "scene characteristics" emerging in visual processing [3].

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2

[2] Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Generative modeling for multi-task visual learning. In International Conference on Machine Learning. PMLR, 2022. 2

[3] H Barrow and J Tenenbaum. Recovering intrinsic scene characteristics. Comput. vis. syst, 1978. 2, 4

[4] 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. 2, 3, 4, 7, 8, 9

[5] Anand Bhattad and D.A. Forsyth. Stylitgan: Prompting stylegan to generate new illumination conditions. In arXiv, 2023. 2

[6] Anand Bhattad and David A Forsyth. Cut-and-paste object insertion by enabling deep image prior for reshading. In 2022 International Conference on 3D Vision (3DV). IEEE, 2022. 3, 7, 8, 9

[7] Anand Bhattad, Daniel McKee, Derek Hoiem, and DA Forsyth. Stylegan knows normal, depth, albedo, and more. In Advances in Neural Information Processing Systems (NeurIPS), 2023. 2

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 4, 1, 5

[9] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. arXiv preprint arXiv:2306.05720, 2023. 2, 1

[10] Katherine Deck and Tobias Bischoff. Easing color shifts in score-based diffusion models. arXiv preprint arXiv:2306.15832, 2023. 4

[11] 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. Ieee, 2009. 3

[12] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2, 3

[13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 2014. 2

[14] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2, 3, 4

[15] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. IEEE transactions on pattern analysis and machine intelligence, 2021. 2, 3

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In The Eleventh International Conference on Learning Representations, 2022. 1, 5

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 2

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010. 2

[19] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. arXiv preprint arXiv:2305.15581, 2023. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020. 2

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. 2

[22] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. arXiv preprint arXiv:2106.05258, 2021. 2

[23] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 2

[24] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 2, 3, 4, 7, 8, 9

[25] 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, 2019. 2

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Proc. NeurIPS, 2020. 4

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 1, 2, 3

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems, 2022. 2

[29] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. arXiv preprint arXiv:2312.02145, 2023. 2

[30] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 2

[31] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891, 2023. 4

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 2

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095, 2022. 1

[34] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In Advances in Neural Information Processing Systems, 2023. 2

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In International Conference on Learning Representations, 2021. 1

[36] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. arXiv preprint arXiv:2307.14331, 2023. 1, 5

[37] Atsuhiro Noguchi and Tatsuya Harada. Rgbd-gan: Unsupervised 3d representation learning from natural image datasets via rgbd image synthesis. In International Conference on Learning Representations, 2020. 2

[38] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In International Conference on Learning Representations, 2021. 2

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 3

[40] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2

[41] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems, 32, 2019. 2

[42] 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, 2022. 1, 2

[43] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 2

[44] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In ACM SIGGRAPH 2022 conference proceedings, 2022. 1, 3, 4

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881, 2023. 2

[46] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. Advances in neural information processing systems, 29, 2016. 2

[47] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In International conference on machine learning. PMLR, 2016. 2

[48] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463, 2019. 4

[49] Pascal Vincent. A connection between score matching and denoising autoencoders. Neural computation, 2011. 2

[50] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3

[51] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. arXiv preprint arXiv:2303.04803, 2023. 2

[52] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. International Journal of Computer Vision, 2021. 2

[53] 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. 3

[54] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 1

[55] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 3

[56] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In International Conference on Learning Representations, 2021. 2

[57] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 2

[58] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. ICCV, 2023. 2

Appendices

A. I-LoRA Pipeline

Fig. 7 illustrates the I-LoRA pipeline applied to Stable Diffusion's UNet in a single-step manner.

B. Additional Ablation Studies

B.1. Rank Efficiency

Our single-step I-LoRA model, distinguished by its high quantitative performance, serves as the basis for ablation studies that assess the influence of rank and labeled data quantity on intrinsic extraction efficiency. We verify that the requirements for compute, parameters, and data to learn I-LoRA are minimal.

Fig. 8 shows surface normal predictions across LoRA ranks. The highest accuracy is achieved with Rank 8, balancing accuracy and memory. Notably, a Rank 2 LoRA with only 0.4M additional parameters (a mere 0.04% increase) still yields good performance. Note that across different generative models, Rank 8 adaptors adds only 0.17% to 0.57% additional parameters (Tab. 2).

B.2. Label Efficiency

The impact of the labeled data size is analyzed in Fig. 9. I-LoRA reaches peak performance using a modest 4000 training examples, with credible predictions visible from as few as 250 samples.

B.3. Number of Diffusion Steps

To assess the impact of the number of diffusion steps on the performance of the multi-step I-LoRA$_{\text{AUG}}$ model, we conducted an ablation study. The results are presented in Fig. 10. For all our experiments in the main text, we used DPMSolver++ [33]. Interestingly, the quality of results did not vary significantly with an increased number of steps, indicating that 10 steps are sufficient for extracting better surface normals from the Stable Diffusion. Nevertheless, we use 25 steps for all our experiments because it is more stable across different image intrinsics.

B.4. CFG scales

When working with the multi-step I-LoRA$_{\text{AUG}}$, the quality of the final output is influenced by the choice of classifier-free guidance (CFG) scales during the inference process. In Fig. 11, we present a comparison of the effects of using different CFG scales. Based on our experiments, we found that using CFG=3.0 results in the best overall quality and minimizes color-shift artifacts.

C. Baselines

C.1. Superiority of I-LoRA over Fine-tuning and Linear Probing

We compare I-LoRA with two common baselines: linear probing and full model fine-tuning. Following Chen et al.[9] for linear probing and employing standard fine-tuning practices, we train all methods with a small dataset of 250 samples to 16000 samples. Our findings, detailed in Tab. 4 and illustrated in Fig. 12, indicate that I-LoRA significantly outperforms these baselines in low-data regimes, validating its superior efficacy and data efficiency.

C.2. Other Ablations and Baselines

We extensively study the effect of applying LoRA to different attention layers within Stable Diffusion models. Specifically, we investigate the outcomes of targeting up-blocks, mid-block, down-blocks, cross-attention, and self-attention layers individually. We find (Fig. 13) that isolating LoRA to up or down blocks or the mid-block alone is less effective or diverges, and applying to either cross- or self-attention layers yields decent results, though combining them is best.

Additionally, we evaluated other image editing methods such as Textual Inversion [16] and VISII [36], alongside InstructPix2Pix's response to "Turn it into a surface normal map" instruction [8]. As shown in Fig. 14, these methods perform poorly for intrinsic image extraction, demonstrating the effectiveness of our I-LoRA approach in extracting scene intrinsics.

C.3. Baseline of Directly Applying SDEdit

In addition to baselines we discussed above, here we show that directly applying SDEdit [35] will also fail to extract reasonable image intrinsics. We take the model from the SDv1-5 column in Fig. 6 of the main paper and apply SDEdit. In Fig. 15, we show directly applying SDEdit results in severe artifacts, regardless of strength.

D. Hyper-parameters

In Table 5, we show the hyperparameters we use for each model.

E. Generated Images Used for Quantitative Analysis

In Tab. 2 of the main paper, we report quantitative results on synthetic images. For Autoregressive models and GANs, we first randomly sample 500 noises and use them to generate 500 RGB images. The same 500 noises will then be used to generate intrinsics with our learned LoRAs loaded. For Stable Diffusion experiments (both single-step and multi-step), we use a single dataset with 1000 synthetic images with

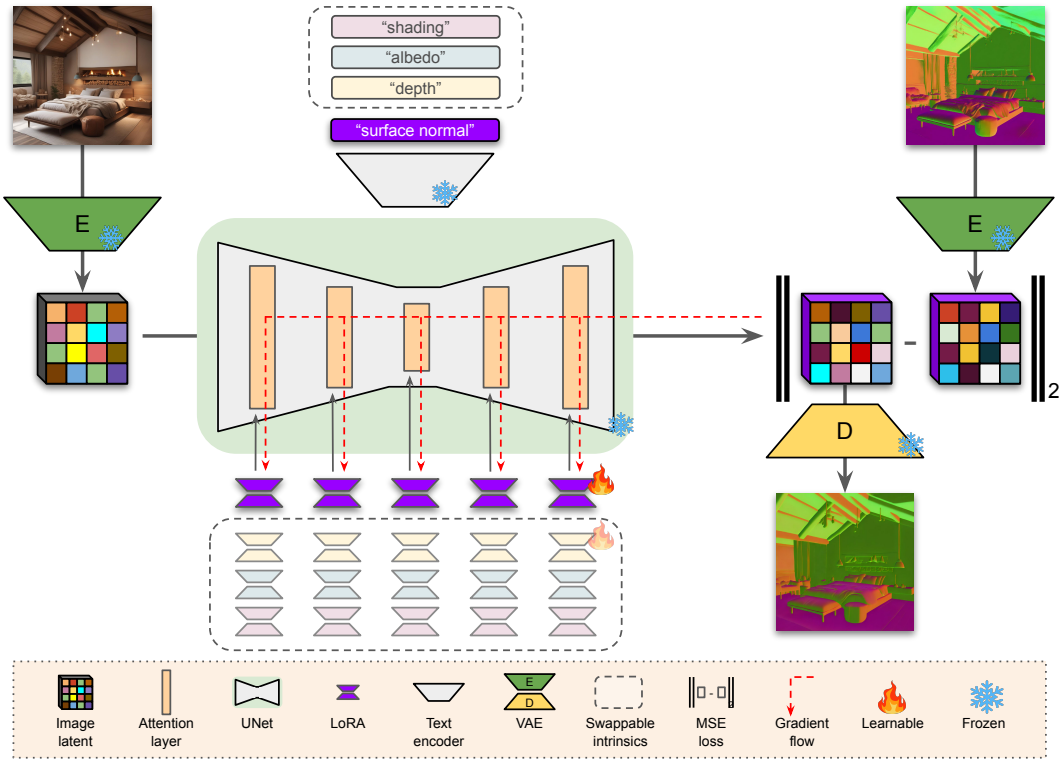
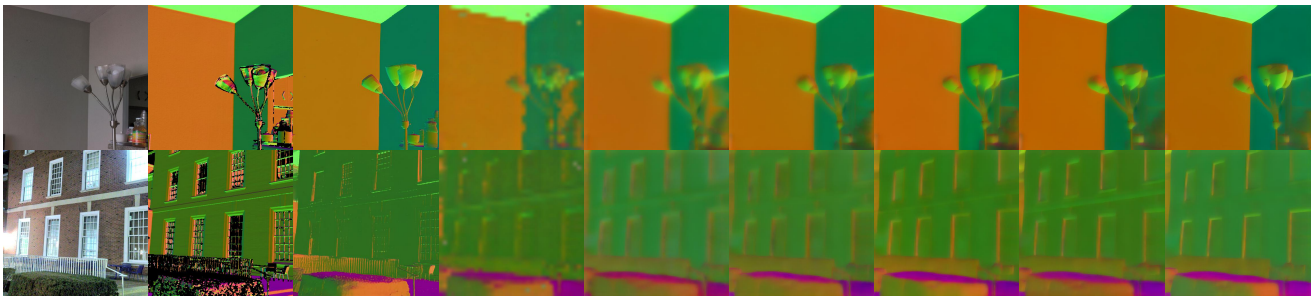


Figure 7. Overview of I-LoRA applied to Stable Diffusion's UNet in a single-step manner. We adopt an efficient fine-tuning approach, specifically low-rank matrices corresponding to key feature maps – attention matrices – to reveal scene intrinsics. Distinct low-rank adaptors (LoRA) are optimized for each intrinsic (***violet*** adaptors for surface normals; swappable with other intrinsics). We use a few labeled examples for this fine-tuning and directly extract scene intrinsics using the same decoder that generates images, circumventing the need for specialized decoders or comprehensive model re-training.

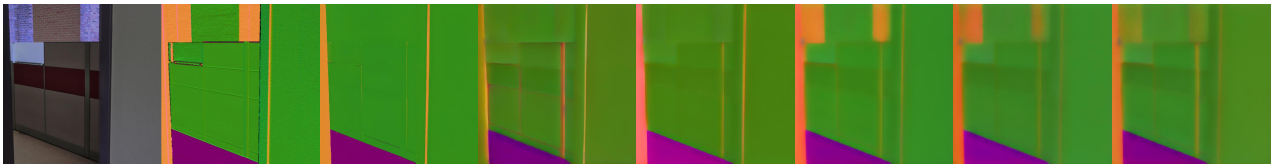


	(a) Real	(b) GT	(c) OD-v2	(d)DINOv2	(e) rank 2	(f) rank 4	(g) rank 8	(h) rank 16	(i) rank 32
Mean Angular Error°↓			18.90	19.74	22.28	22.57	20.31	21.17	21.84
L1 Error (× 100)↓			15.21	16.00	18.14	18.39	16.53	17.19	17.81
LoRA Param.			-	0.26%	0.04%	0.08%	0.17%	0.34%	0.68%

Figure 8. Parameter Efficiency of I-LoRA. We evaluate I-LoRA across various rank settings for surface normal extraction. Lower ranks such as 8 offer a balance between efficiency and effectiveness. All model variants are trained using SD's UNet (v1.5) with 4000 samples. Performance metrics, such as Mean Angular Error and L1 Error for normals, and additional parameter counts are detailed below each variant.

Table 4. We find LoRA to consistently outperform all baselines for different number training samples (first row).

	250		1000		4000		16000	
	Mean Error°↓	L1 $\times_{100}$ ↓	Mean Error°↓	L1 $\times_{100}$ ↓	Mean Error°↓	L1 $\times_{100}$ ↓	Mean Error°↓	L1 $\times_{100}$ ↓
Linear Probe	29.10	23.74	28.45	23.25	28.52	23.26	28.22	23.11
Fine-tuning	34.40	27.58	25.19	20.28	28.03	22.17	27.39	22.24
LoRA (Ours)	**27.73**	**22.46**	**22.22**	**18.05**	**20.31**	**16.53**	**21.26**	**17.33**



	(a) Real	(b) GT	(c) OD-v2	(d) 250	(e) 1000	(f) 4000	(g) 16000	(h) 24895
Mean Angular Error°↓			18.90	27.73	22.22	20.31	21.26	21.64
L1 Error (× 100) ↓			15.21	22.46	18.05	16.53	17.33	17.64

Figure 9. Data efficiency of I-LoRA. We report results from varying training samples. Even with 250 samples, I-LoRA captures surface normals. We observe the best performance with 4k samples. Models (d)-(h) all use the same SD UNet(v1-5) and rank 8 LoRA. Note: SOTA supervised model (c), was trained using 12M+ labeled training samples.

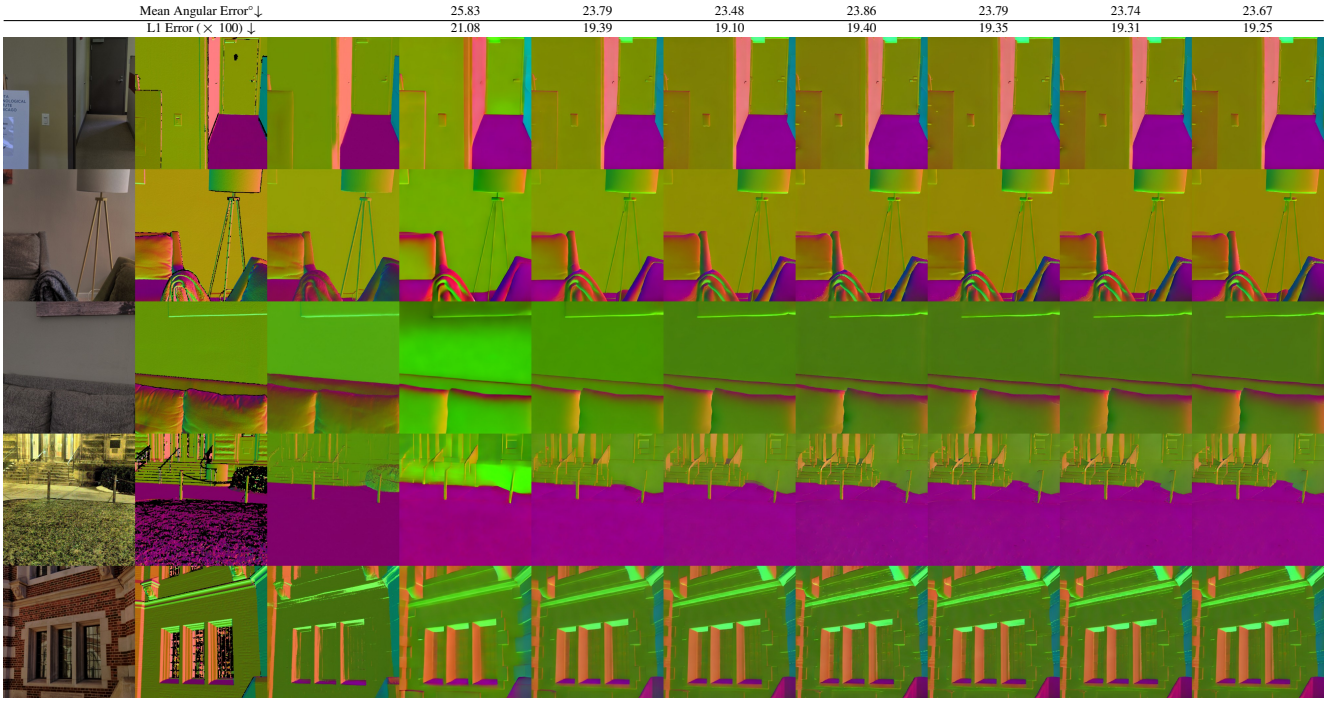


Figure 10. Ablation study to determine the effect of varying numbers of diffusion steps while keeping CFG fixed at 3.0. Our findings show that there are very small differences, both in terms of quantity and quality, after 10 steps. For our main paper, we report results for 25 steps as it is more stable across different intrinsics.

various prompts. The pseudo GT are obtained by applying SOTA off-the-shelf models on the RGB images.

F. Additional Qualitative Results

In Fig. 16, we present comparisons for I-LoRA$_{\text{AUG}}$ and I-LoRA1-5$_{\text{AUG}}$. Fig. 17 and Fig. 18 shows extra results for models trained on FFHQ dataset. More examples of scene intrinsics extracted from StyleGAN-v2 trained on LSUN bedroom can be found in Fig. 19. In Fig. 20, we show results for Stable Diffusion I-LoRA (single-step) on generated images. Shown in Fig. 21 are extra results for StyleGAN-XL trained on ImageNet.

G. Results on 1024^2 synthetic images

Our multi-step I-LoRA$_{\text{AUG}}$ models, although trained exclusively on 512^2 images from the DIODE dataset, demonstrate their robustness by successfully extracting intrinsic images from 1024^2 high-resolution synthetic images generated by Stable Diffusion XL [39], as shown across Figures 22 to 31

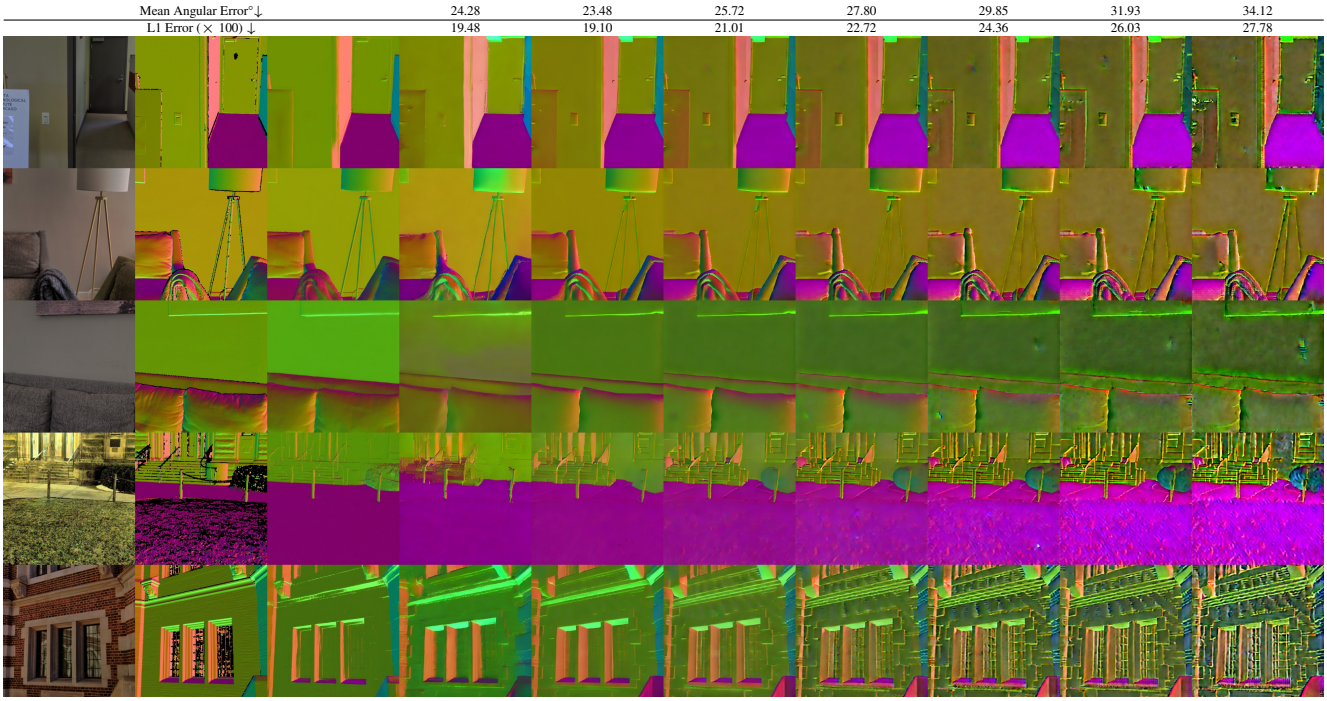


Figure 11. Ablation study analyzing the impact of different classifier-free guidance (CFG) on I-LoRA$_{\text{AUG}}$ surface normal prediction. For efficiency, we experimented with a step of 10. We observed that CFG=1 sometimes led to incorrect semantic predictions, particularly in the case of stairs in row 4. On the other hand, using large CFGs (5 and beyond) results in more severe color shift problems.

Figure 12. Comparison with baselines. All models are trained with 250 samples. Note LoRA effectively extracts better normals compared to other baselines.

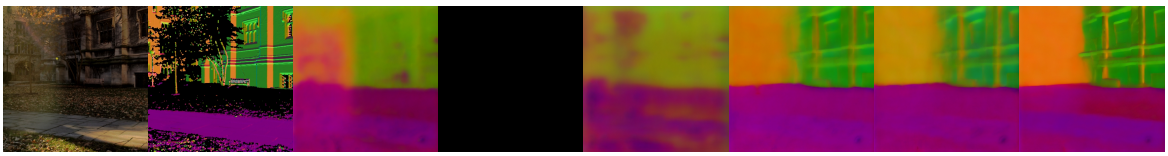


Figure 13. Ablation study on the effect of applying LoRA on different types of attention layers. We started all models with SD v1-5, 4000 training samples and LoRA rank=8.

Figure 14. Comparison of image editing techniques for surface normal mapping. VISII and Textual Inversion yield unsatisfactory results, while InstructPix2Pix fails to interpret the task, resulting in near-original output.

Figure 15. We observe applying SDEdit method on the SDv1-5 model alone, without incorporating the additional input image latent encoding, fails to produce satisfactorily aligned and high-quality scene intrinsics. The reason for this might be the considerable domain shift that exists between RGB images and surface normal maps, which results in severe artifacts when using SDEdit. The variable "s" represents the strength of SDEdit.

Model	Dataset	Resolution	Rank	LR	BS	LoRA Params	Generator Params	Convergence Steps
VQGAN	FFHQ	256	8	1e-03	1	0.13M	873.9M	$\sim$ 4000
StyleGAN-v2	FFHQ	256	8	1e-03	1	0.14M	24.8M	$\sim$ 4000
StyleGAN-v2	LSUN Bedroom	256	8	1e-03	1	0.14M	24.8M	$\sim$ 4000
StyleGAN-XL	FFHQ	256	8	1e-03	1	0.19M	67.9M	$\sim$ 4000
StyleGAN-XL	ImageNet	256	8	1e-03	1	0.19M	67.9M	$\sim$ 4000
I-LoRA$_{\text{AUG}}$ (multi step)	Open	512	8	1e-04	4	1.59M	943.2M	$\sim$ 30000
I-LoRA (single step)	Open	512	8	1e-04	4	1.59M	943.2M	$\sim$ 15000

Table 5. Hyper-parameters for each model. LR refers to the learning rate and BS refers to the batch size. Please note that the number of steps required to reach convergence reported above is for normal/depth. However, it is worth noting that albedo and shading tend to require significantly fewer steps to converge (usually half of normal/depth). Additionally, I-LoRA$_{\text{AUG}}$ (multi-step) and I-LoRA (single-step) are trained on real-world DIODE dataset, while the other models are trained on synthetic images within a specific domain. (Num. of params of VQGAN counts transformer + first stage models; Num. of params of I-LoRA$_{\text{AUG}}$ and I-LoRA counts VAE+UNet)

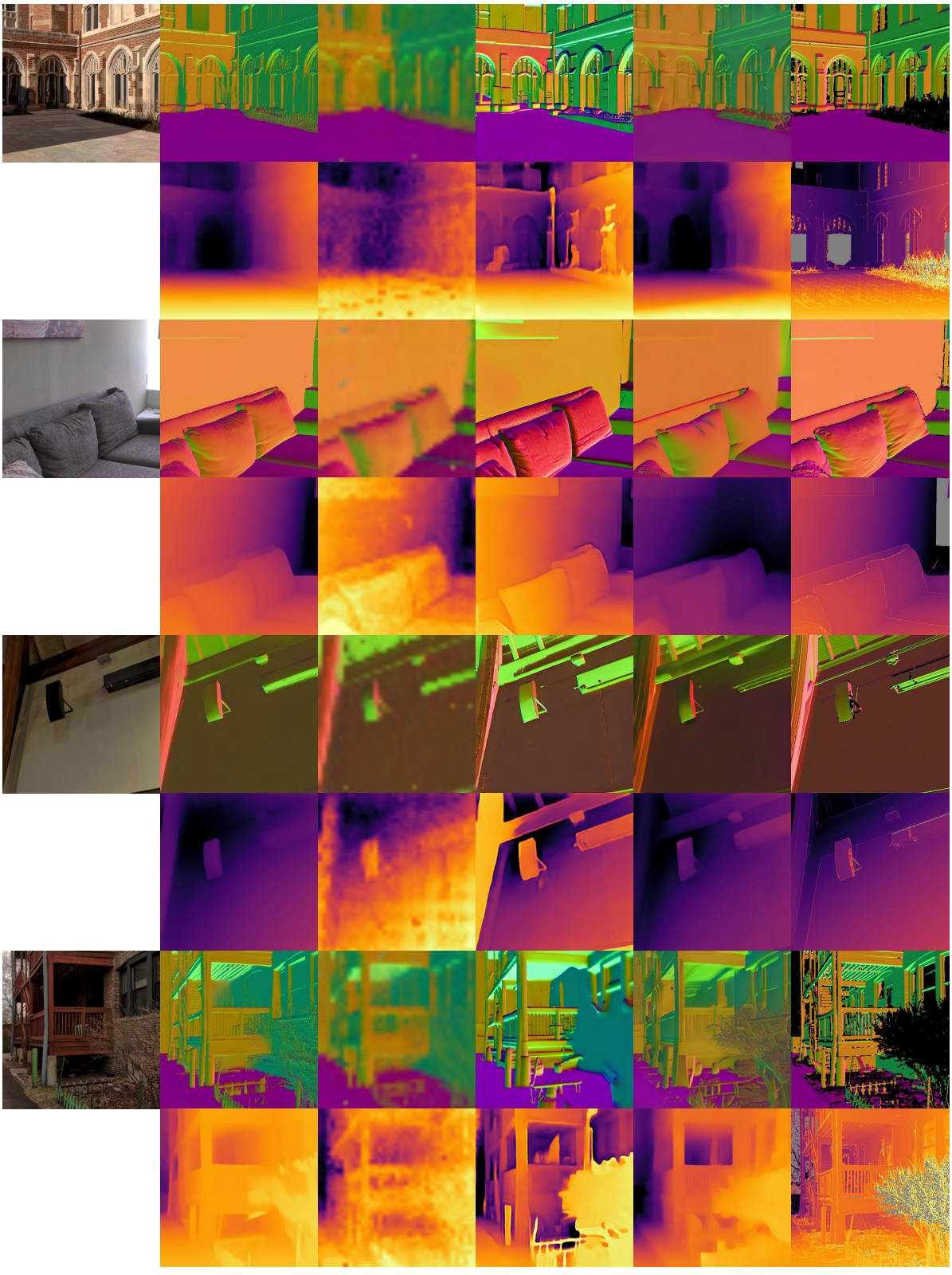


Figure 16. Additional results after applying improved diffusion techniques with I-LoRA$_{AUG}$. I-LoRA$_{AUG}$ was found to significantly reduce color shift artifacts observed in I-LoRA1-5$_{AUG}$ during the extraction of detailed scene intrinsic results.

Figure 17. Scene intrinsics from different generators – VQGAN, StyleGAN-v2, and StyleGAN-XL – trained on FFHQ dataset: The "image" column shows the synthetic images produced by each model. Subsequent columns show four scene intrinsics extracted by a SOTA non-generative model and I-LoRA(ours).

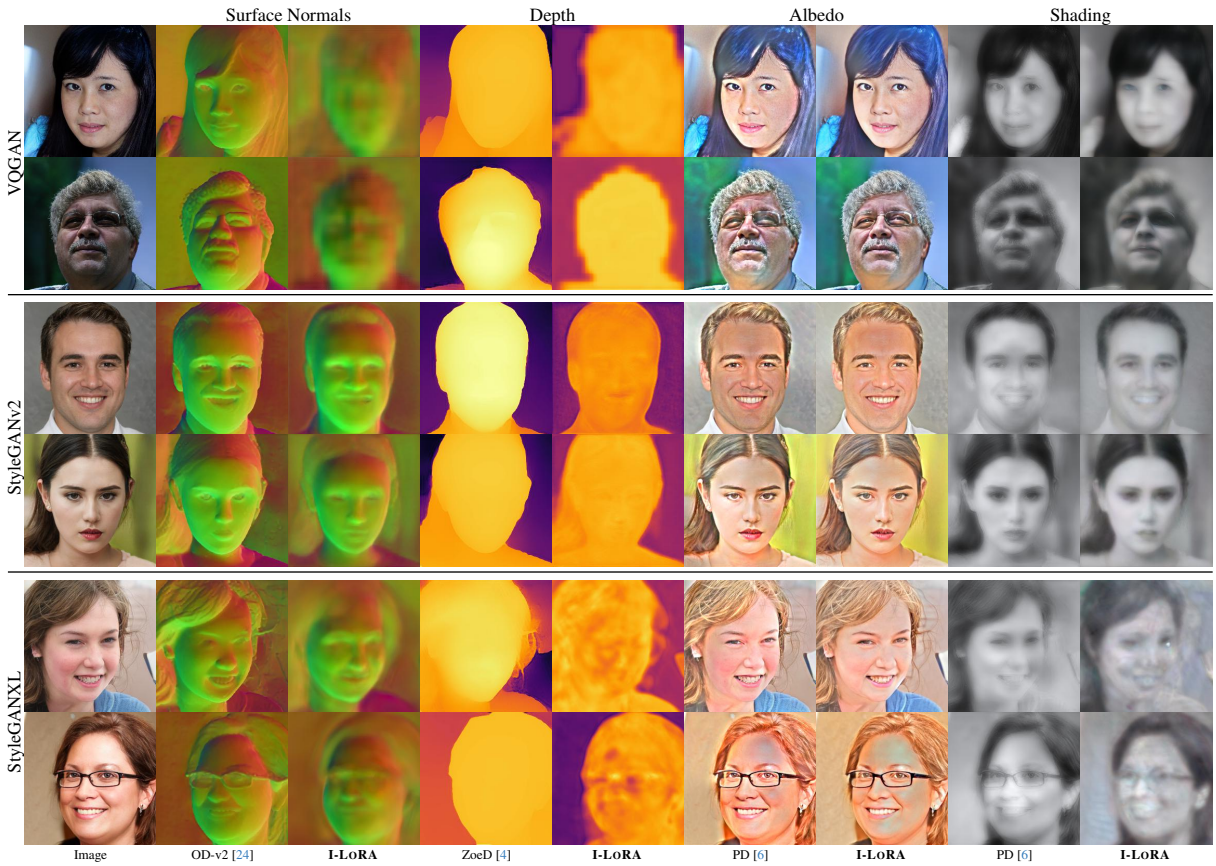


Figure 18. Additional results of scene intrinsics from different generators – VQGAN, StyleGAN-v2, and StyleGAN-XL – trained on FFHQ dataset.

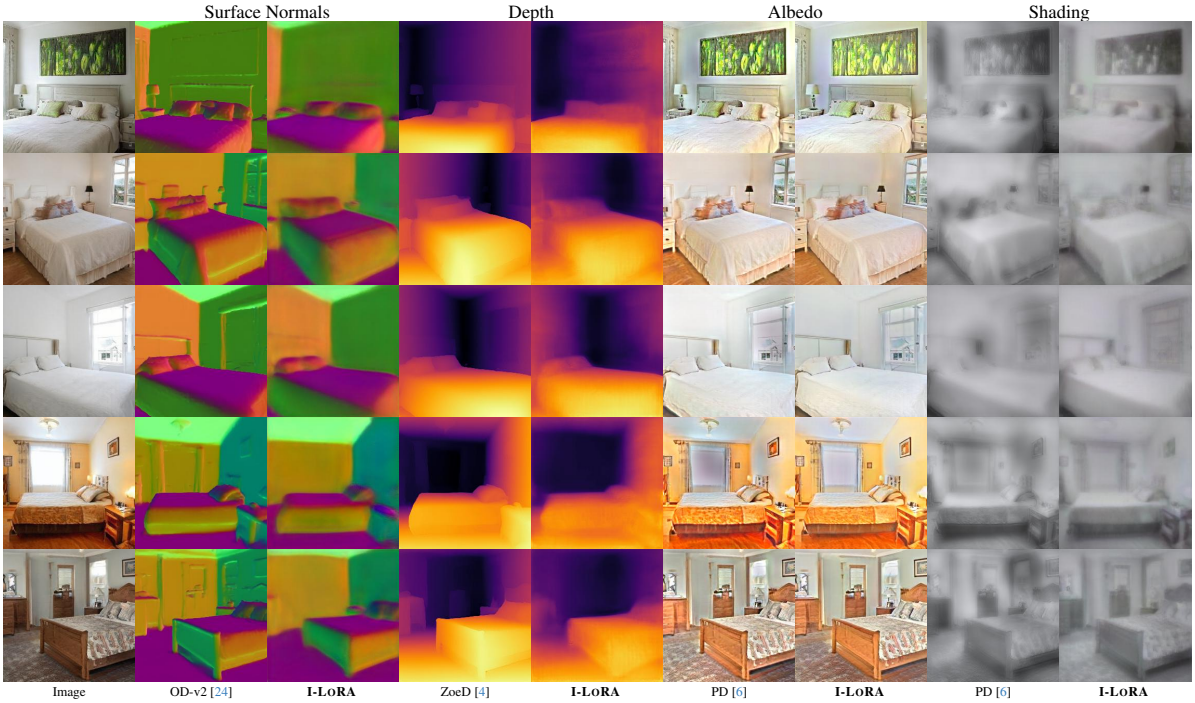


Figure 19. Additional results of scene intrinsics extraction from Stylegan-v2 trained on LSUN bedroom images.

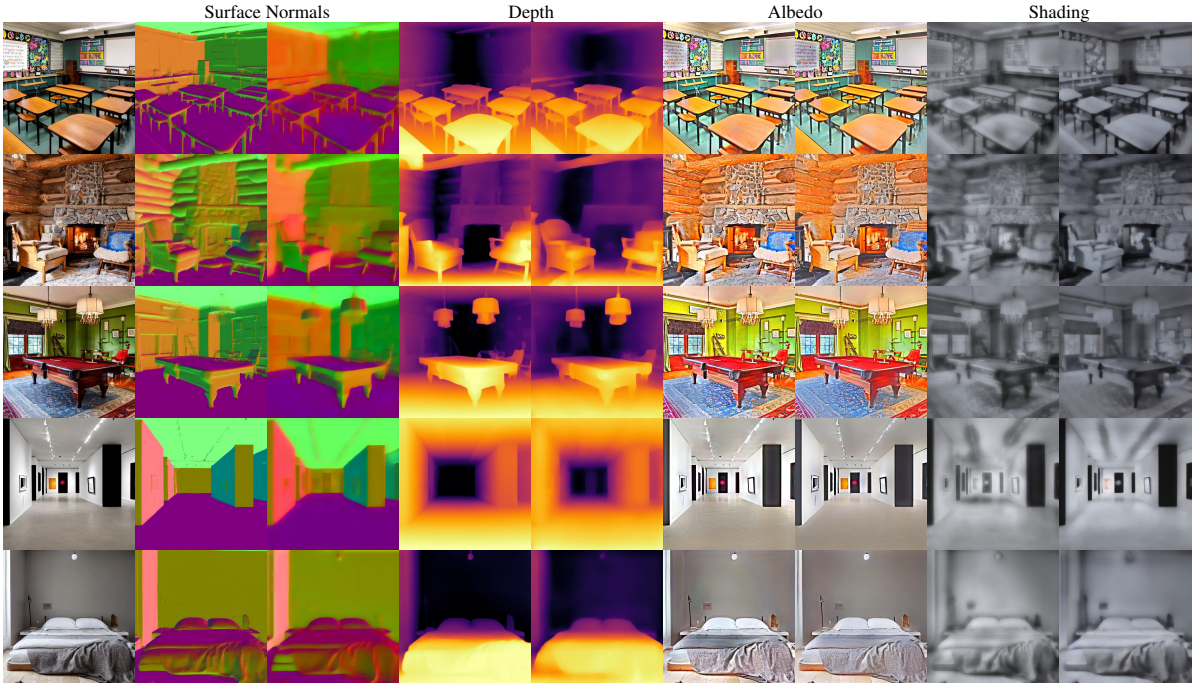


Figure 20. Additional results of scene intrinsics extraction from Stable Diffusion I-LoRA (single-step).

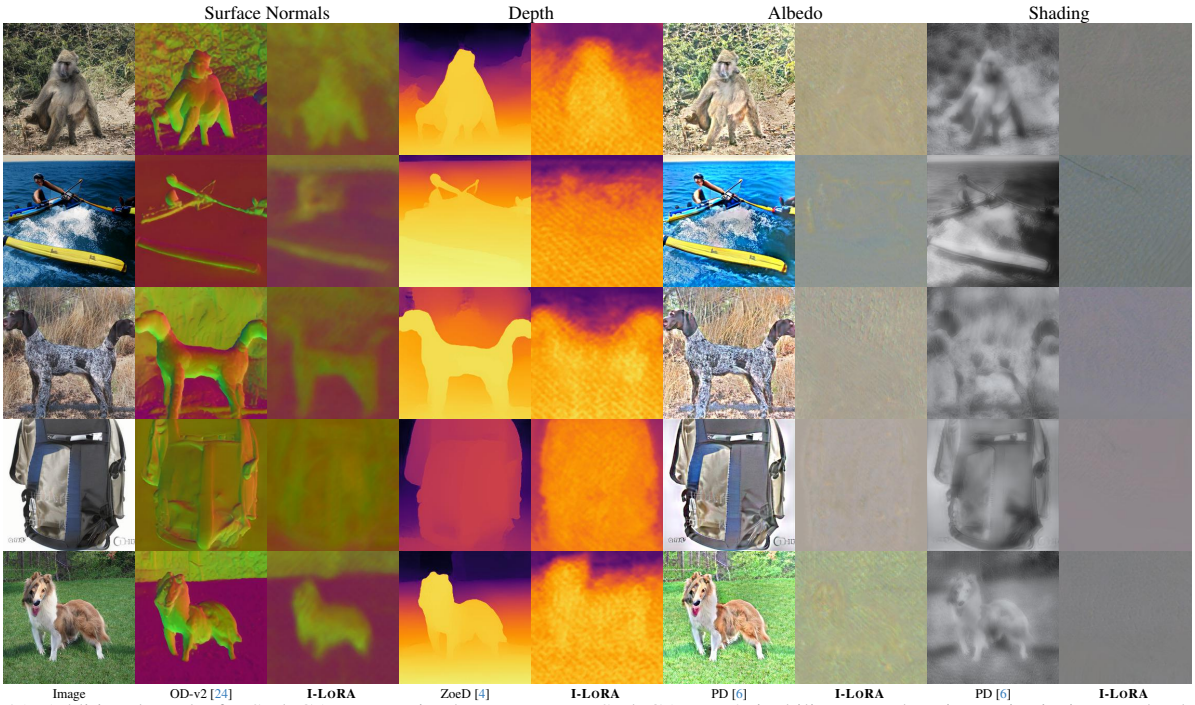


Figure 21. Additional results for StyleGAN-XL trained on ImageNet. StyleGAN-XL's inability to produce image intrinsics may be due to its inability to create high-quality plausible images.

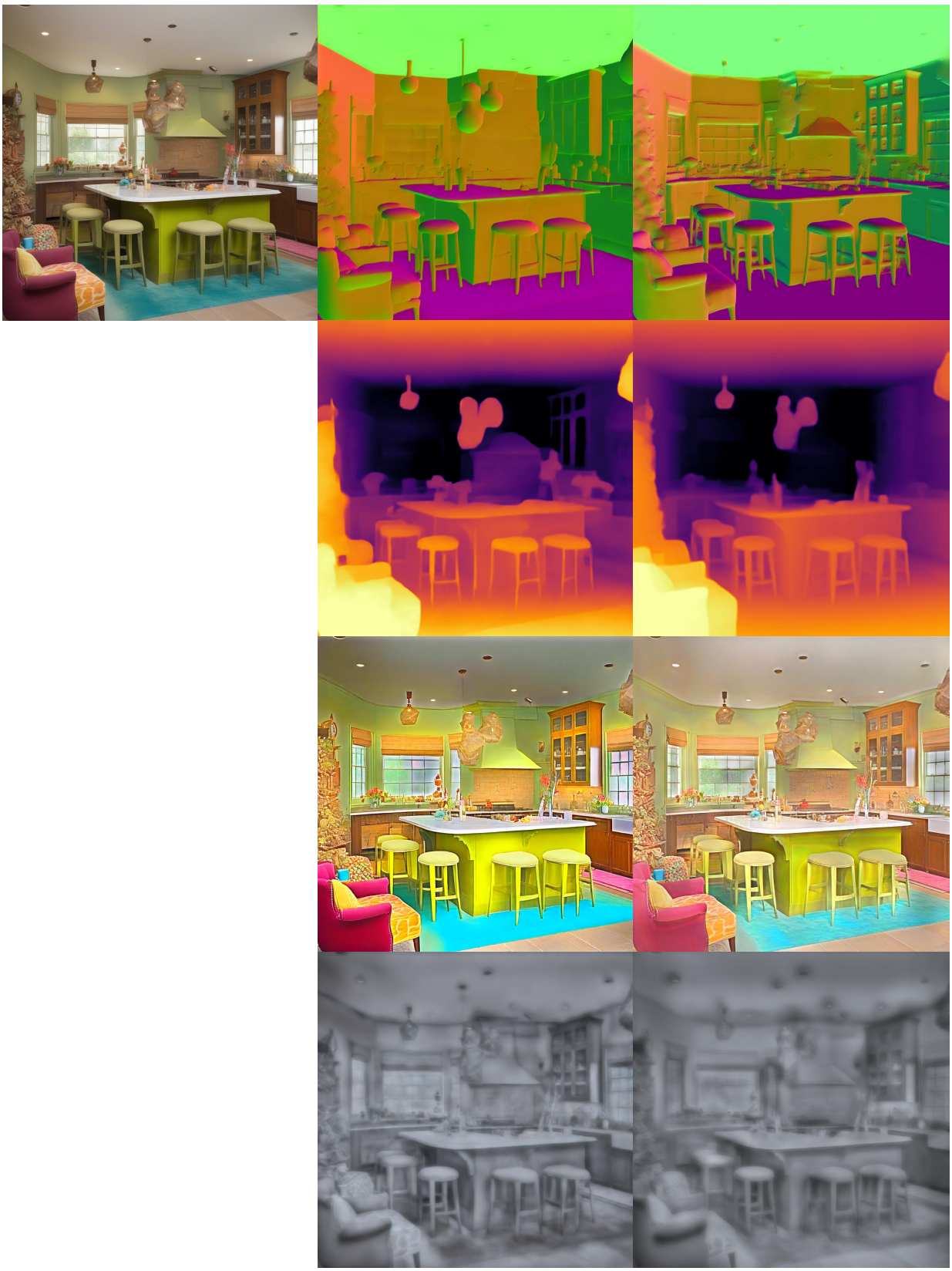


Figure 22. Results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

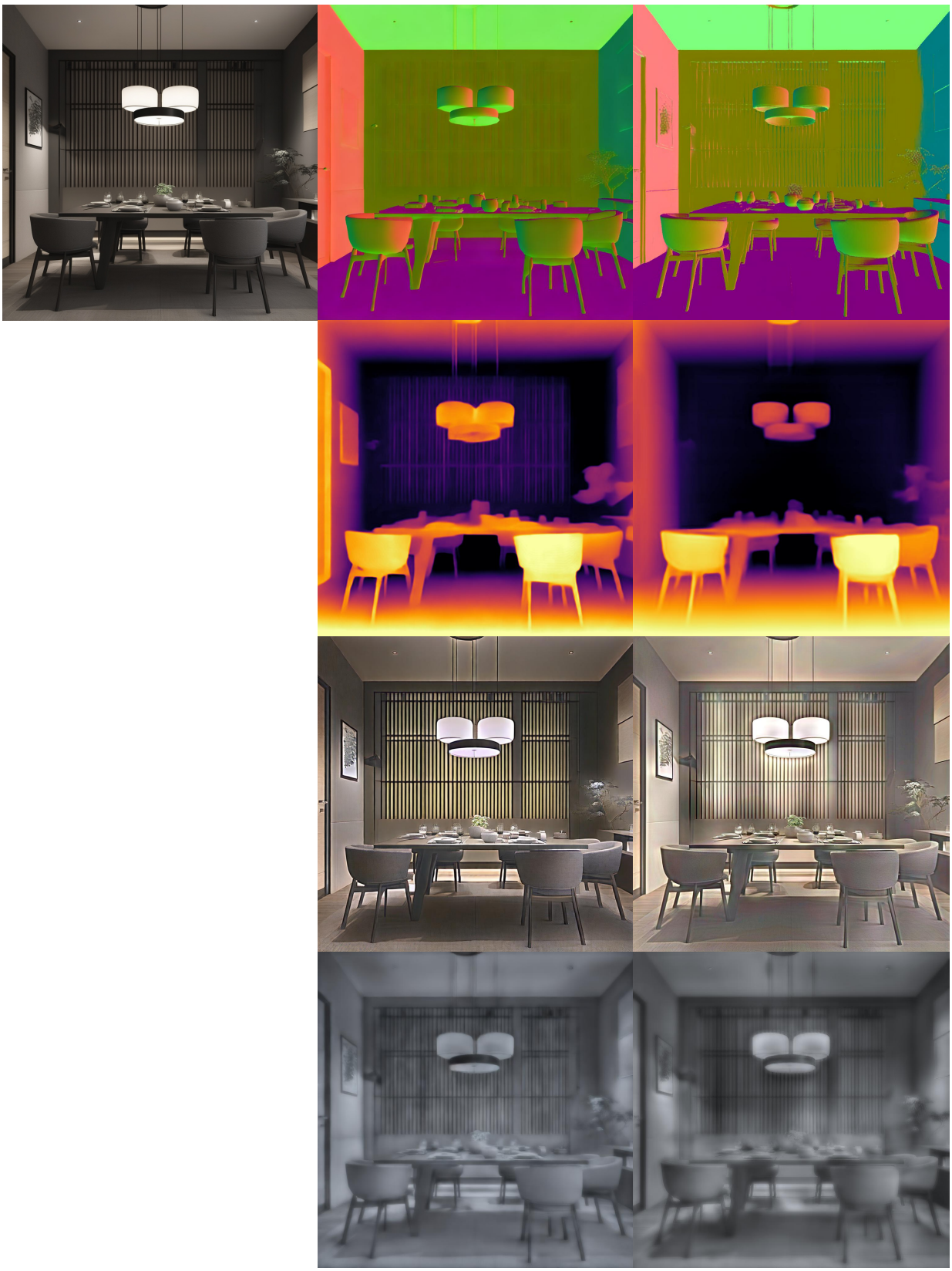


Figure 23. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

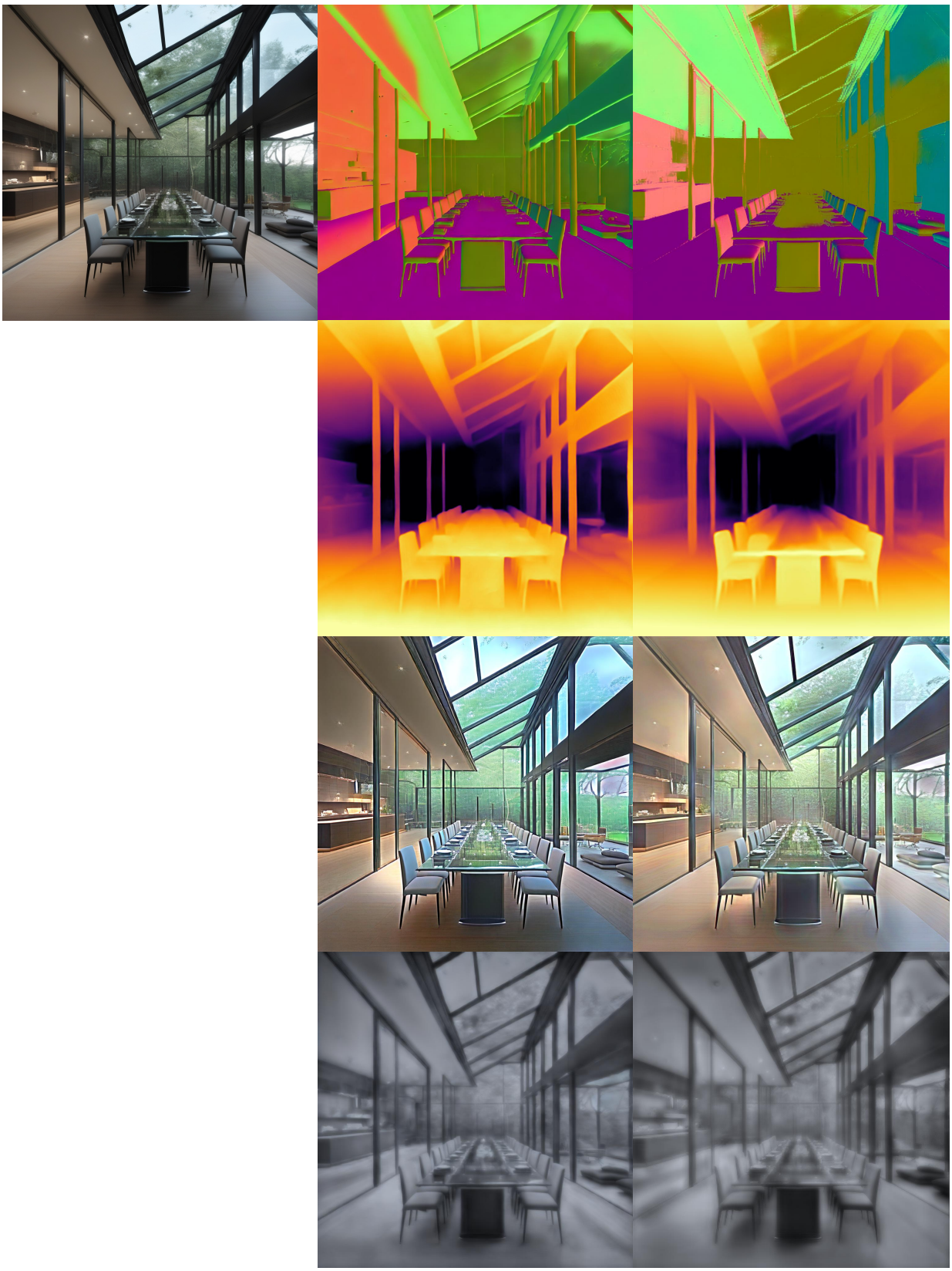


Figure 24. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

Figure 25. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

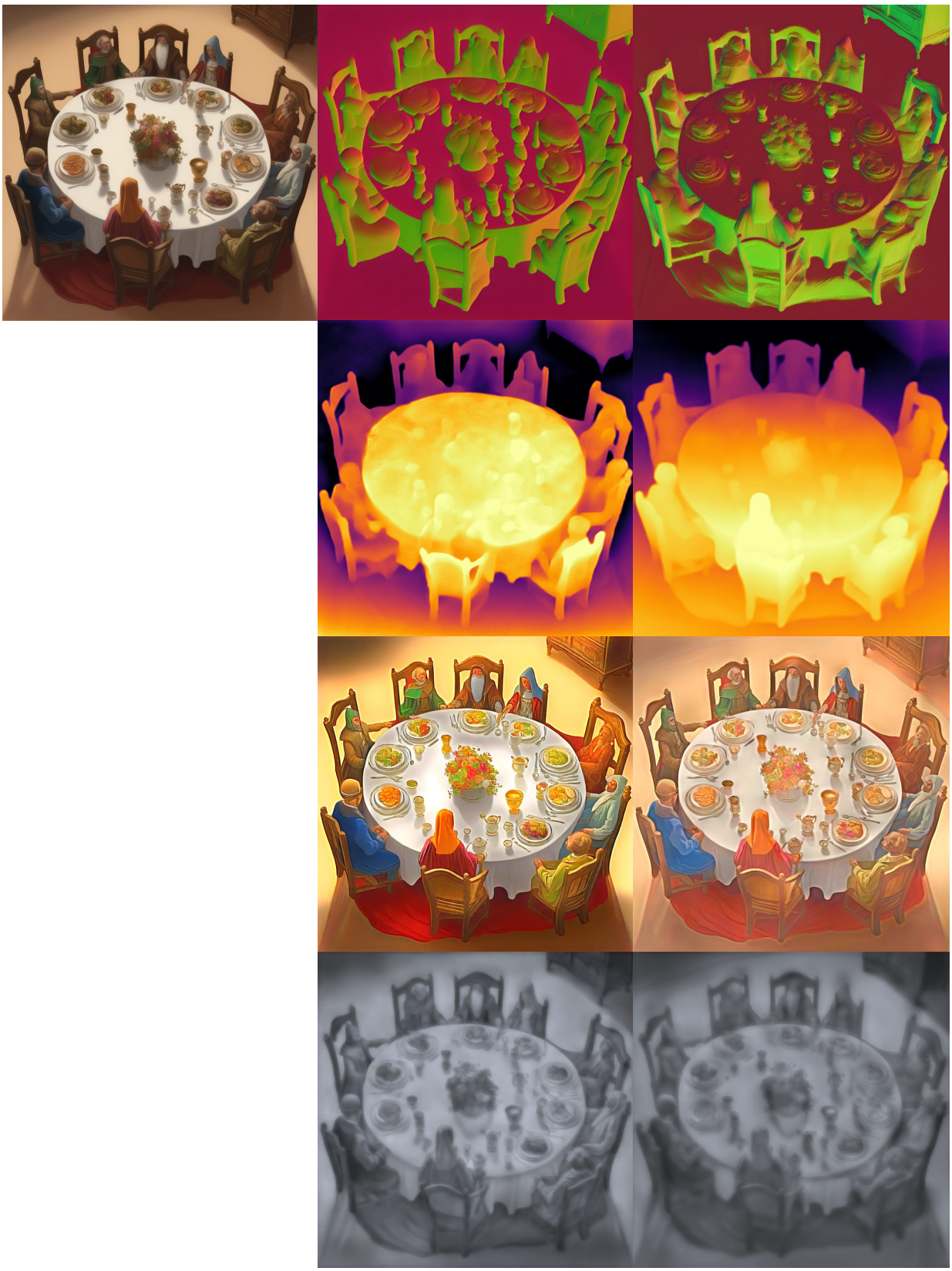


Figure 26. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

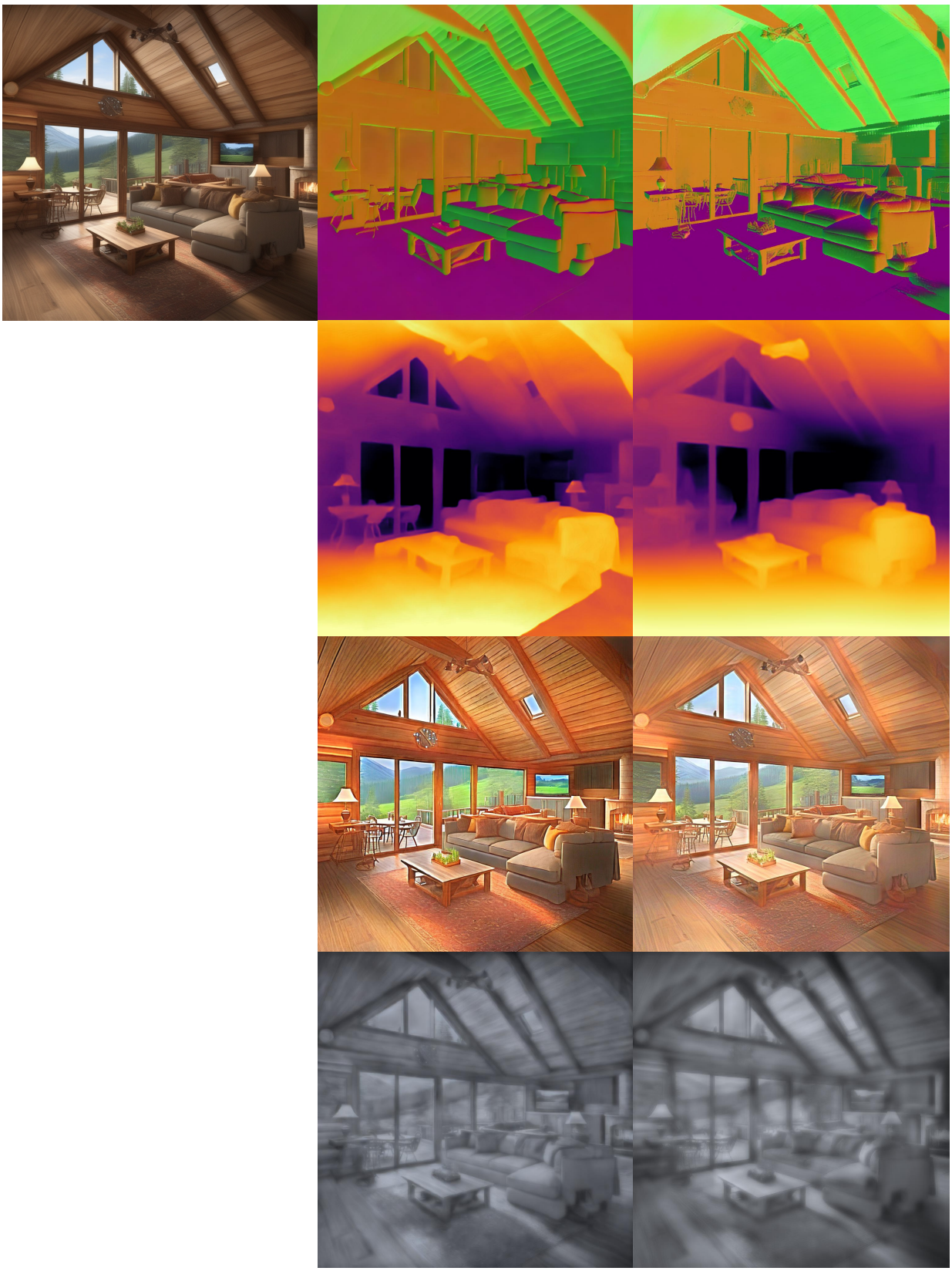


Figure 27. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

Figure 28. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

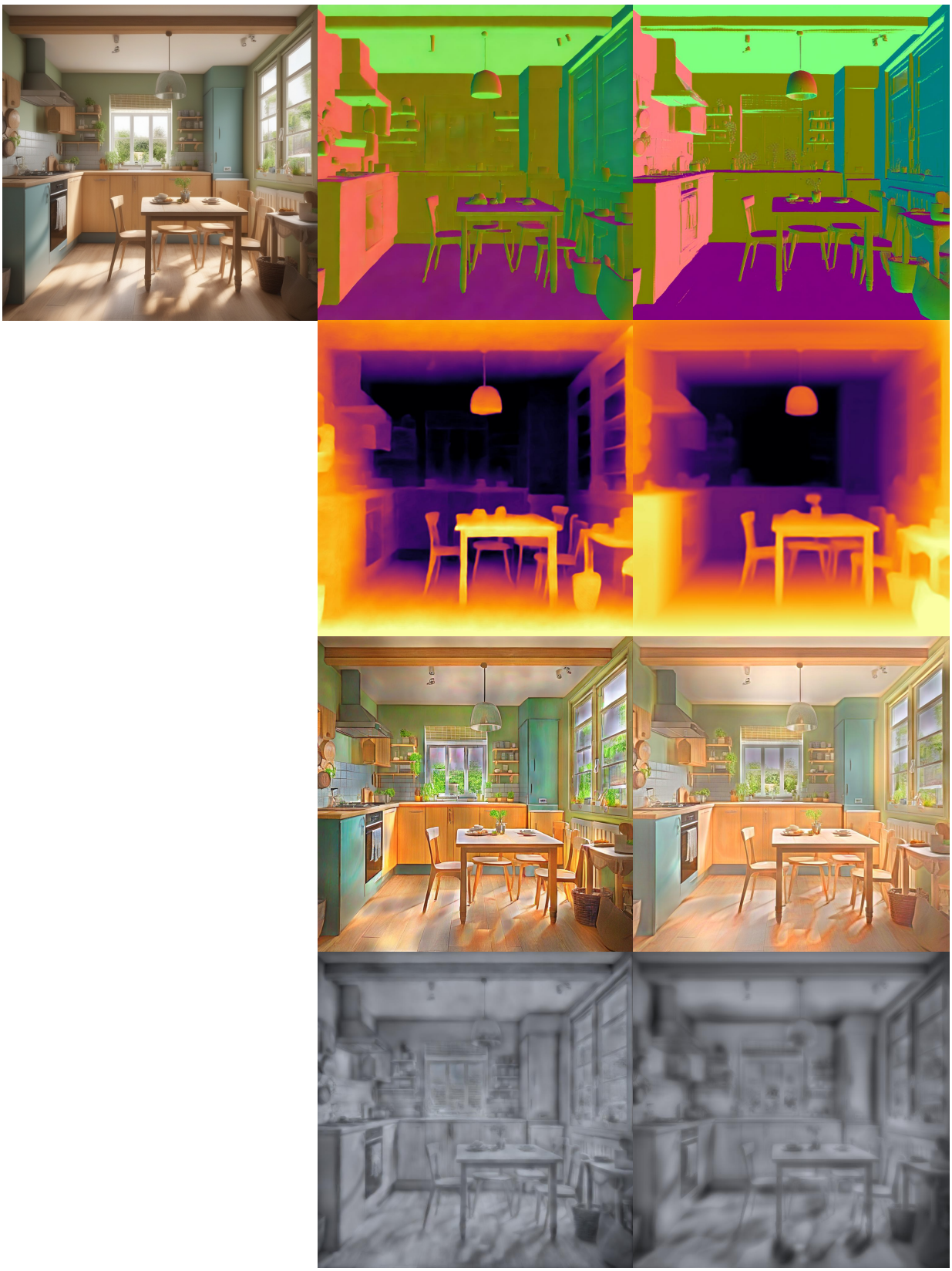


Figure 29. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

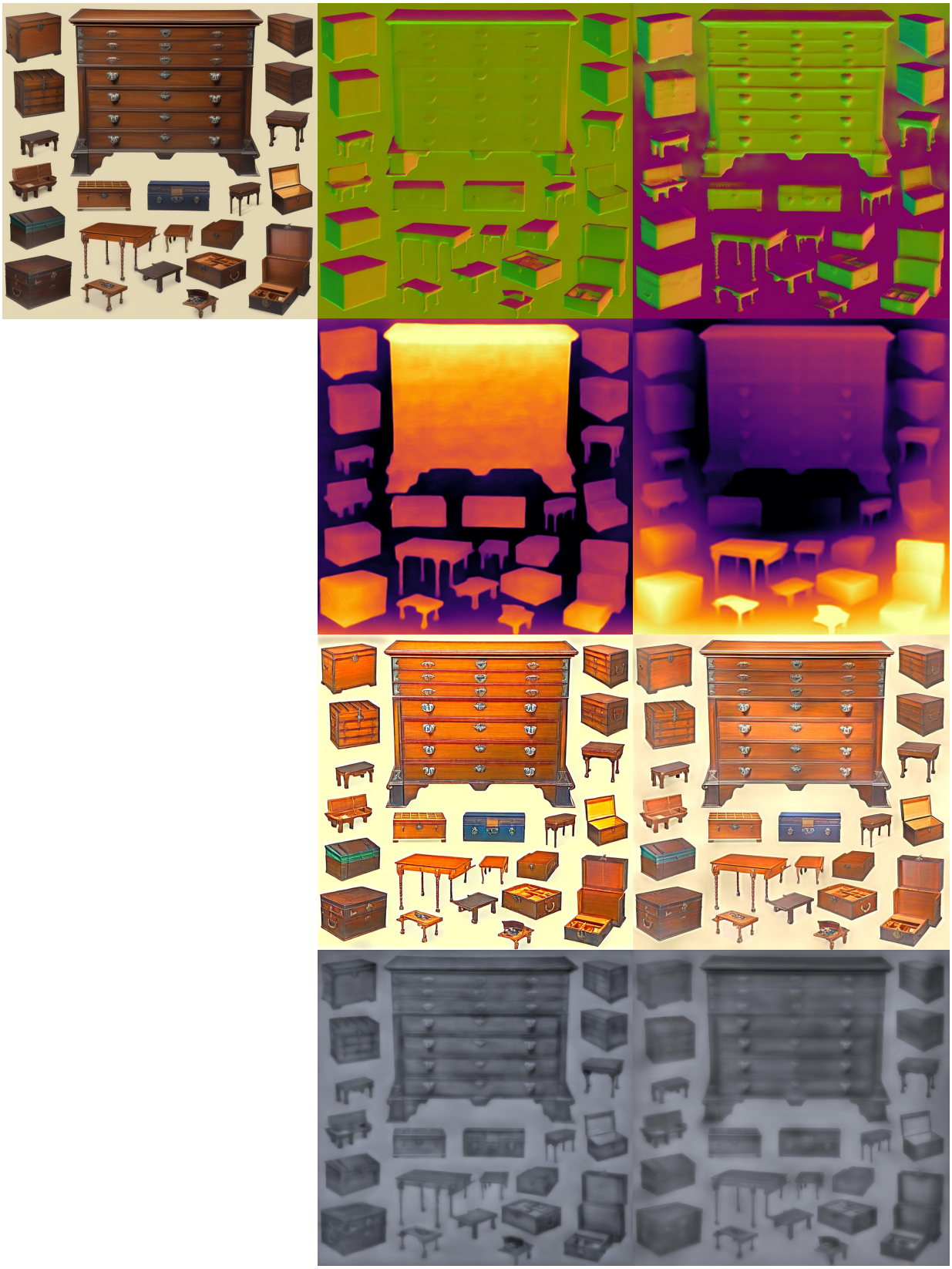


Figure 30. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.

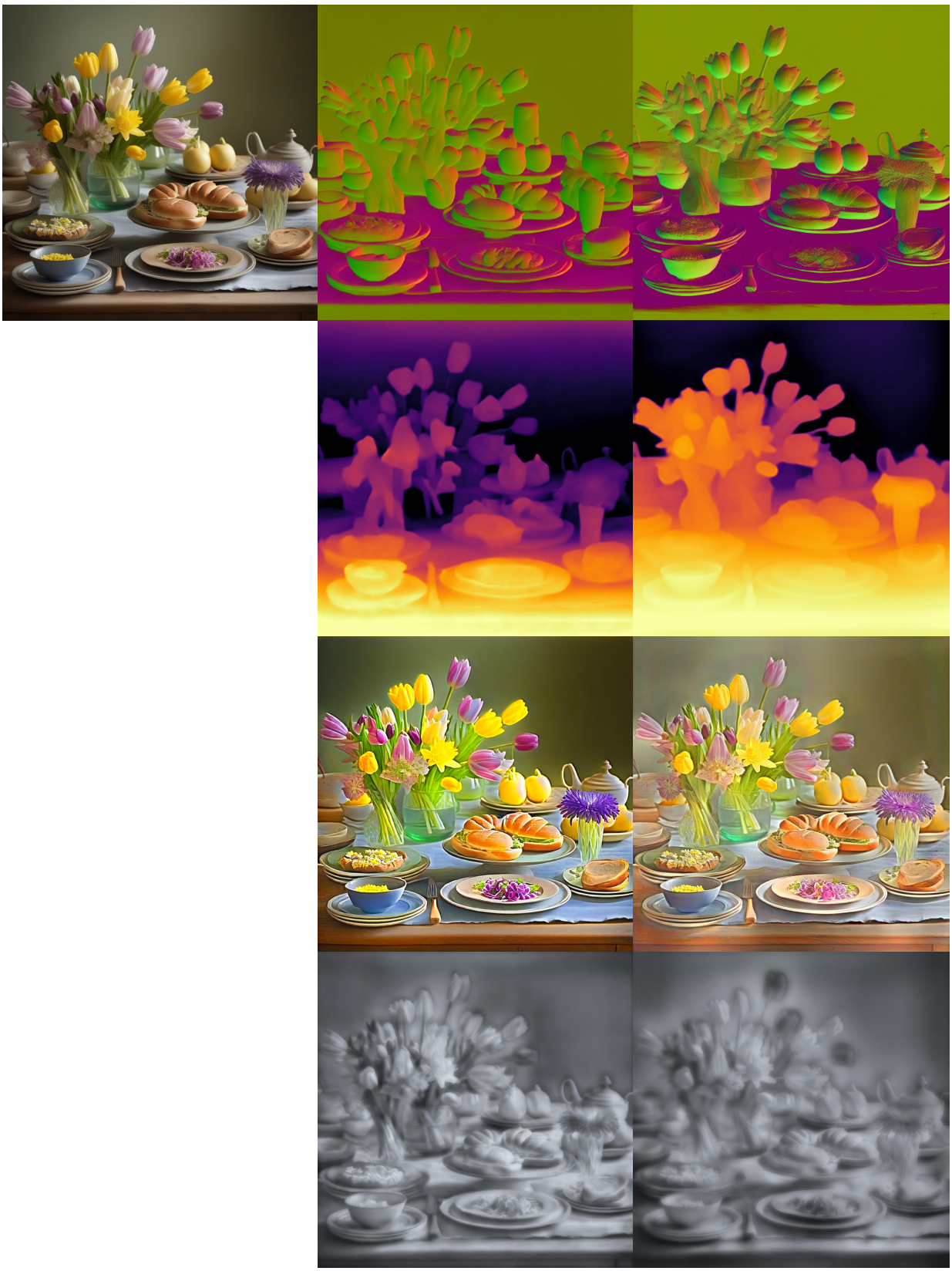


Figure 31. Cont. results of I-LoRA$_{\text{AUG}}$ models applied on unseen 1024^2 synthetic images. Left: original image; middle: ours; right: pseudo ground truth.