Supplementary

Quantitative results

Table S1 contains an extended listing of the numbers in Tab. 1. Specifically, we additionally report the distribution matching scores on the individual PBR components (albedo, roughness, metallic, and normal bump maps) as well as the distribution matching scores on three additional mixed three-channel maps:

- **PBR_1:** grayscale albedo, roughness, metallic
- **PBR_2:** grayscale albedo, normal bump map X, normal bump map Y
- **PBR_3:** roughness, metallic, $\|\text{normal bump XY}\|_2$

Following Chambon et al. [5], we average the sum of all the constituent PBR distribution matching scores to reflect the total distribution matching score, as the underlying neural networks are limited to three-channel images. We only use the bump map X and Y values: as a bump map value of $[0, 0, 1]^T$ indicates no surface change the bump map only has two degrees of freedom.

Qualitative results

Please refer the later pages of this supplementary material for an extended qualitative comparison between the ablated variants. For each of the 29 objects selected for the OOD test set, we have randomly selected one out of the 5 prompts and visualize the results for one random seed for each object-prompt combination on a single page. Finally, we also show a single random seed for the four remaining prompts for the proposed method in the very last pages.

Extended ablation study discussion

Communication direction. One-way communication from the frozen RGB model to the PBR model is not successful; see Eq. (1) for a discussion why the RGB model's output distribution needs to be aligned to the PBR model. Clockwise communication (RGB encoder → PBR encoder, PBR decoder → RGB decoder) *is* successful, but contrary to that work full bi-directional communication still improves the results slightly across the board.

Communication type. The pixel-wise MLP slightly outperforms the pixel-wise single-layer zero-convolution accross the board. In the qualitative comparison later in this supplementary material, we note that there is very little visual difference between both options, so that we settle for the simpler one with less parameters. The attention-based communication performs surprisingly well despite the lack of spatial correspondences: it is able to somewhat match the distributions, and

the resulting PBR images are visually acceptable — although noticeably less well aligned than those from the pixel-wise communication methods.

Collaborative Control vs. Fine-tuning. Both approaches seem to work reasonably well in terms of the distribution matching metrics, however (referring to Tab. 1), the fine-tuning approaches are slightly less visually pleasing according to QAlign [72]. Yet in the qualitative comparison it becomes clear that the fine-tuning approaches perform significantly worse; they are far less detailed, and often completely fail to represent the prompt.

PBR VAE vs. RGB VAE on triplets. Already from the quantitative numbers it becomes clear that the PBR VAE is significantly better than the RGB VAE. This is also reflected in the qualitative results, where the PBR VAE is able to generate much more realistic PBR images than the RGB VAE. From the quantitative results, performance on the albedo maps appears slightly better (expectedly so, as it is encoded in its separate triplet, and is roughly in-domain for the pre-trained VAE) than the other PBR components, although it too is not modeled well in comparison to the other variants.

Training budget. Quantitatively, at least, it is clear that the increased training budget improves continues to improve the performance. This is harder to see visually, and anecdotally we have found that the output quality converges relatively soon, but that the capacity to reflect the prompt continues improving during the longer training regimen in addition to be aided by the octupled batch size (12 vs 96).

Training resolution. Oddly, the quantitative numbers here do not reflect the qualitative results. The higher resolution training seems to have a significant impact on the visual quality of the output, but the quantitative numbers do not reflect this (although there is a slight indication in this direction in the QAlign numbers for the OOD prompts). We attribute this to two factors. First, that both the CLIP embedding from the CLIP-MMD score and the Inceptionv3 embedding from the FID score are relatively high-level embeddings, which don't reflect low-level details well. Secondly, that the training dataset in Objaverse contains many materials for which the textures have no high-frequency details, and that the higher resolution training is therefore not as beneficial as it would be for a more diverse dataset — which is reflected in our OOD prompts that eke out more spatial detail.

Data sparsity. Both quantitatively and qualitatively, it is clear that removing the text prompt cross-attention layer from the PBR model is crucial; the distribution matching scores rise significantly even when the entire training dataset is leveraged, which is only exacerbated in more data-sparse regimes. The behavior for more data-sparse regimes *without* this cross-attention is a bit more mixed, and it is hard to draw a conclusion from the quantitative numbers alone. The qualitative results are slightly more informative, showing that even the variant trained on only 1% of the dataset (≈60,000 images) is able to generate visually acceptable PBR images, although results improve slightly as more data becomes available.

		2% held-out evaluation data																	
		CMMD ↓									FID ↓								
		Albedo	Roughness	Metallic	Bumps	PBR_1	PBR_2	PBR_3	mean(PBR)	Relit	Albedo	Roughness	Metallic	Bumps	PBR_1	PBR_2	PBR_3	mean(PBR)	Relit
Communication	one-way	27.22	15.41	24.17	9.39	13.41	13.85	11.65	16.44	13.38	28.70	21.15	23.23	15.76	20.23	21.67	15.55	20.90	16.39
	clockwise	5.83	9.85	16.15	4.62	3.41	2.65	4.96	6.78	2.76	16.68	12.56	18.53	7.64	10.73	11.08	8.28	12.21	11.53
	bi-directional	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	Pixel-wise zero-conv	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	Pixel-wise MLP	4.11	8.89	10.83	4.92	2.68	2.11	4.46	5.43	1.87	15.41	12.65	15.87	7.58	10.04	10.46	8.00	11.43	10.67
	Global Attention	11.46	8.71	11.57	3.40	6.31	6.42	5.33	7.60	5.22	21.20	12.86	17.05	8.29	13.25	13.56	9.09	13.61	11.93
	Collaborative Control	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	Fine-tuning (with RGB output)	5.34	9.27	54.39	9.20	4.62	2.37	8.65	13.40	2.78	15.81	12.41	34.40	8.10	10.82	10.66	8.73	14.42	10.79
	Fine-tuning (without RGB output)	5.23	6.53	9.08	5.58	3.39	2.01	4.94	5.25	2.88	15.78	11.84	14.41	8.06	10.61	10.87	8.34	11.41	11.37
	PBR VAE	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	RGB VAE on triplets	24.27	73.58	284.07	75.78	44.44	26.68	63.80	84.66	5.99	17.43	27.24	85.12	13.91	12.84	13.47	10.69	25.81	11.63
Training budget	**1 A100, two days**	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	8 A100s, three days	2.95	4.98	3.95	3.08	1.76	1.54	2.46	2.96	1.12	13.55	9.21	12.80	6.34	8.93	9.25	6.78	9.55	9.76
Training resolution	256×256	3.66	3.91	0.91	1.66	1.74	1.70	2.02	2.23	1.44	15.28	9.08	11.81	6.14	9.46	10.11	6.84	9.82	10.20
	512×512	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
Training data	**No PBR prompt attention** 1%	4.10	8.92	13.64	4.54	4.29	3.46	4.81	6.25	1.43	16.18	12.38	15.52	7.88	11.26	11.34	8.51	11.87	10.79
	No PBR prompt attention 5%	3.62	9.17	13.51	4.04	3.07	2.18	4.79	5.77	1.45	15.41	12.54	16.58	7.14	10.21	10.52	8.02	11.49	10.54
	No PBR prompt attention 20%	3.94	9.05	14.64	4.57	2.81	2.22	4.54	5.97	1.68	15.26	12.00	17.57	7.37	10.04	10.48	7.80	11.50	10.61
	No PBR prompt attention 98%	4.28	9.68	15.48	4.74	2.85	2.21	4.85	6.30	1.79	15.63	12.85	16.62	7.55	10.23	10.63	8.07	11.65	10.64
	PBR prompt attention 1%	11.25	24.06	56.65	11.44	12.96	10.76	17.19	20.61	4.25	21.36	22.03	32.84	11.34	14.80	13.43	12.68	18.35	12.16
	PBR prompt attention 5%	6.13	13.42	38.46	6.80	6.68	4.88	8.83	12.17	2.58	17.44	15.46	28.67	9.40	12.07	11.80	9.83	14.95	10.97
	PBR prompt attention 20%	5.18	11.68	37.87	7.26	5.66	4.12	7.70	11.35	2.33	16.38	14.48	30.74	9.05	11.66	11.60	9.57	14.78	10.78
	PBR prompt attention 98%	4.92	12.23	24.09	7.01	5.10	3.70	7.24	9.18	2.57	16.43	13.38	22.65	8.78	11.07	11.38	9.05	13.25	11.02

Table S1: Detailed distribution matching scores for all evaluated variants in the ablation study. The ablation baseline is highlighted in bold, duplicated for easier comparisons within the individual ablations.