Teamwork: Collaborative Diffusion with Low-rank Coordination and Adaptation – Supplemental

SAM SARTOR, College of William & Mary, USA
PIETER PEERS, College of William & Mary, USA

ACM Reference Format:
Sam Sartor and Pieter Peers. 2025. Teamwork: Collaborative Diffusion with Low-rank Coordination and Adaptation – Supplemental. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3757377.3763870

1 Datasets

In Table 1 we show the different components used to train Teamwork. Note there is some uncertainty in how to handle diffuse color in CGIntrinsics. Since that dataset has purely diffuse surfaces, we opted to use the color as summed reflectance so as to not bias the model towards marking usually metallic objects as diffuse.

Some models (e.g., RGB→X) and datasets (e.g., InteriorVerse) opt to use the albedo+metalness workflow instead of the specular+diffuse workflow. In those cases we convert using the following equations (after gamma correction):

$$D = \max(A - S_b, 0) * (1 - M)$$
$$S = (A - S_b) * M + S_b,$$

where $S_b = 0.04$ is the assumed specularity of dielectric surfaces. We employ a gamma of 2.2 for all albedo and shading maps and convert when appropriate.

Most synthetic datasets have occasional flipped normals. We detect and correct these by comparing the normal vector to the camera view vector.

2 Performance

Table 2 quantitatively compared detailed memory usage and computation costs for both inference and training for various methods and base-models. The fully-trained models evaluated in the main paper differ in resolution, rank, and implementation details. To eliminate confounding variables we provide a direct comparison of the architectures themselves, implemented as similarly as possible, with a 1024 resolution, rank of 16, and cast to BF16, without text-encoders, and for a single diffusion step. FLOPs are measured with TORCH.UTILS.FLOP_COUNTER. Close matches to evaluated models are marked with an ⋆ for intrinsic image decomposition, † for SVBRDF estimation, and ‡ for inpainting.

Authors' Contact Information: Sam Sartor, College of William & Mary, Williamsburg, USA, slsartor@wm.edu; Pieter Peers, College of William & Mary, Williamsburg, USA, ppeers@siggraph.org.

SA Conference Papers '25, December 15–18, 2025, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763870

Table 1. Summary of components in intrinsic decomposition and material estimation datasets used to train Teamwork.

	InteriorVerse	HyperSim	CGIntrinsics	Infinigen	MatFusion	ABC Renders
Diffuse Albedo	✓	✓	✗	✓	✓	✓
Specular Albedo	✓	✗	✗	✗	✓	✓
Summed Albedo	✓	✗	✓	✗	✓	✓
Roughness	✓	✗	✗	✗	✓	✓
Normals	✓	✓	✗	✓	✓	✓
Depth	✓	✓	✗	✓	✗	✓
Diffuse Shading	✗	✓	✗	✓	✗	✓
Shading	✓	✗	✓	✗	✓	✓
Specular Residual	✗	✓	✗	✗	✗	✓

For Joint Attention we use Flash Attention which has an optimized linear memory usage with respect to the number of tokens. However, Flash Attention retains the quadratic computation cost of naive attention. When evaluating Flux, TORCH.UTILS.FLOP_COUNTER fails due to high memory requirements. Finally, note that the number of trainable parameters for RGB→X is higher than all other methods because it requires full fine-tuning.

3 SVBRDF Results

Below are are all 50 materials from the MatFusion test set, evaluated on the published MatFusion colocated-lighting model and on each of our teamwork models. Teamwork operates on the upscaled 512×512 maps, while MatFusion operates at the original 256 × 256.

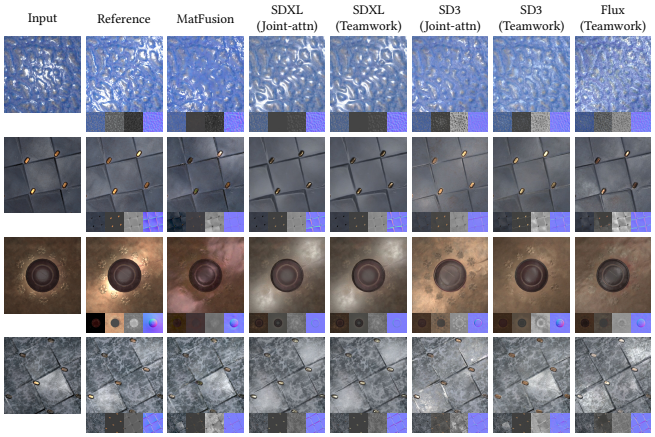
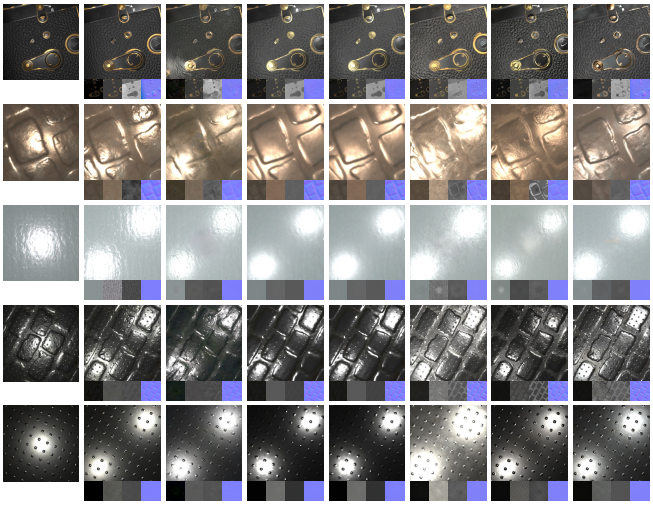
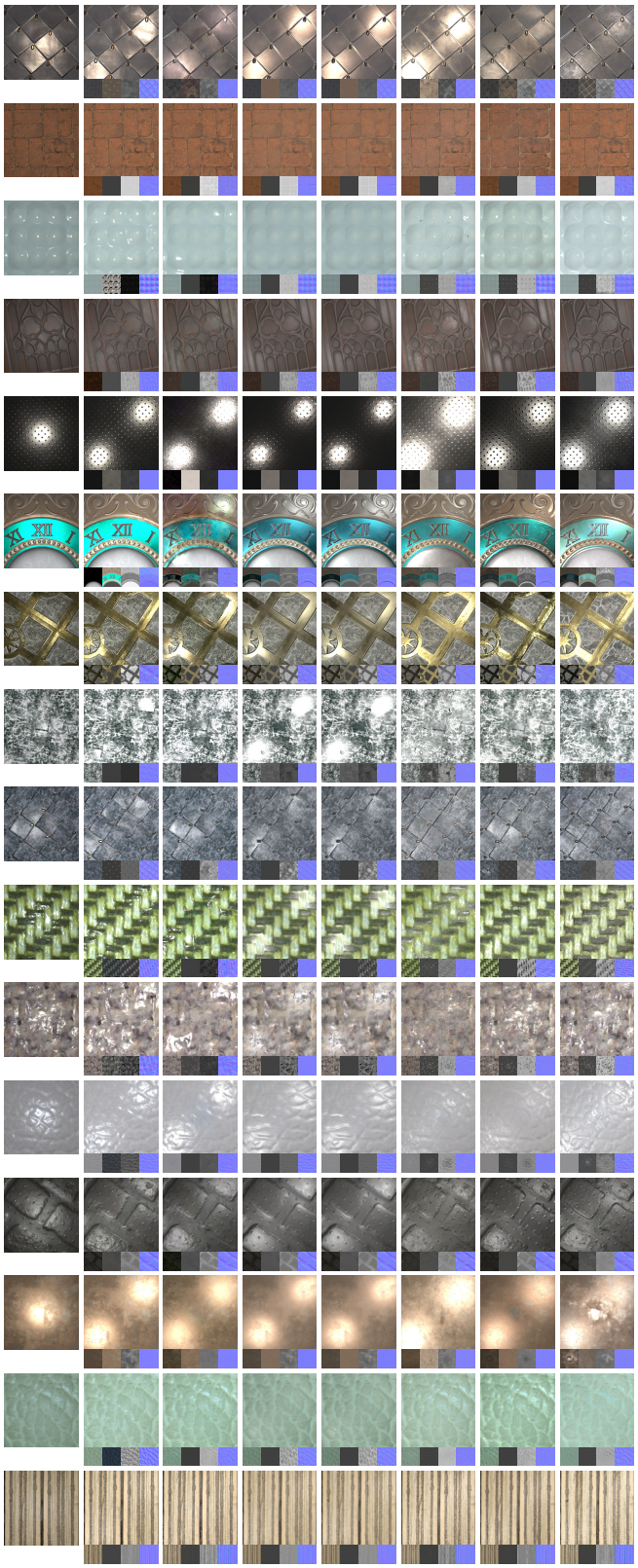
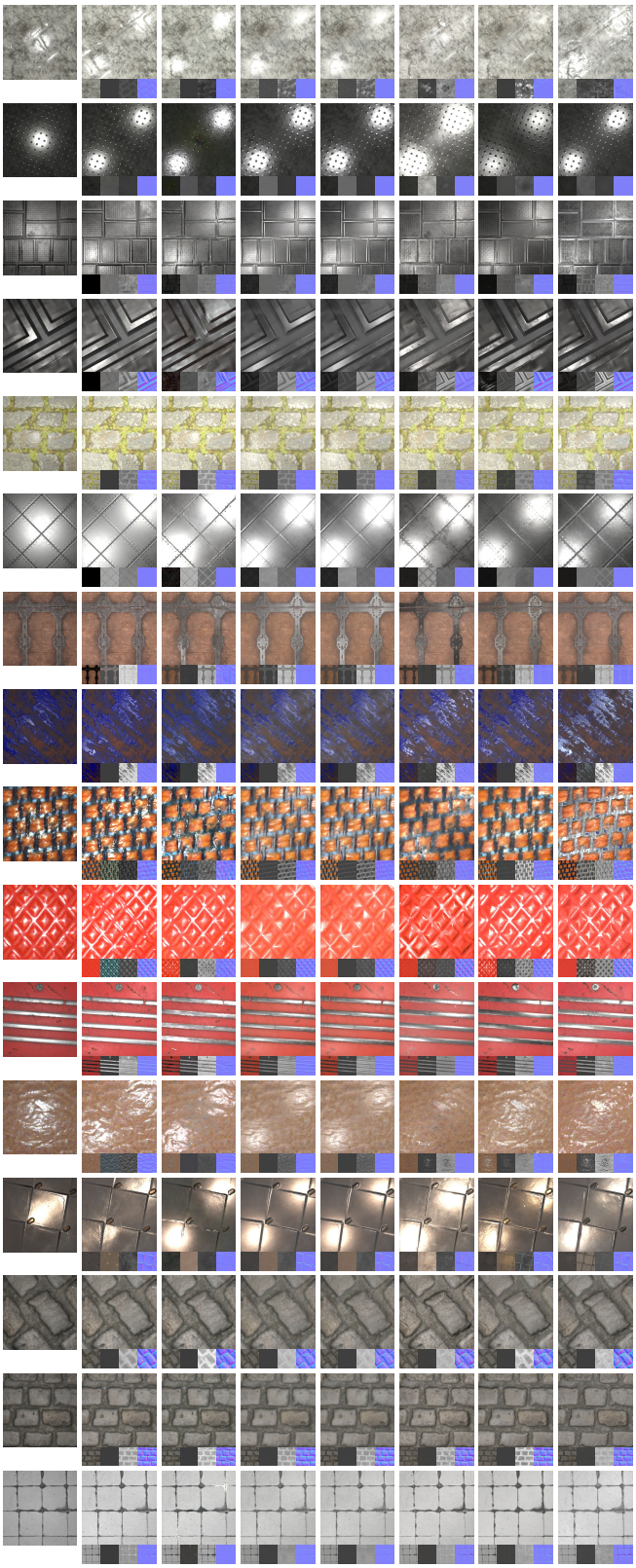


Table 2. Quantitative comparison of measured computation and memory costs for different techniques.

T	Total #	Trainable #	Infer VRAM	Train VRAM	Infer FLOPs	Train FLOPs	
SD2 —RGB→X							
3	865.9M	865.9M	2.4GiB	9.3GiB	9.4T	37.7T	
6	865.9M	865.9M	3.1GiB	11.2GiB	23.4T	94.3T	★
8	865.9M	865.9M	3.6GiB	13.0GiB	32.7T	132.0T	
10	865.9M	865.9M	4.1GiB	14.8GiB	42.1T	169.7T	
SDXL —Teamwork							
3	2.7B	132.5M	5.8GiB	8.0GiB	20.6T	64.1T	
6	2.8B	264.9M	6.6GiB	10.8GiB	41.2T	128.3T	†
8	2.9B	353.2M	7.0GiB	12.7GiB	54.9T	171.0T	
10	3.0B	441.5M	7.5GiB	14.6GiB	68.6T	213.8T	
SDXL —Joint Attention							
3	2.7B	132.5M	5.8GiB	8.0GiB	25.3T	85.3T	
6	2.8B	264.9M	6.6GiB	10.8GiB	64.7T	234.1T	†
8	2.9B	353.2M	7.0GiB	12.7GiB	98.8T	368.6T	
10	3.0B	441.5M	7.5GiB	14.6GiB	139.2T	531.3T	
SD3 —RGB→X							
3	2.0B	2.0B	4.5GiB	19.8GiB	17.8T	74.1T	
6	2.0B	2.0B	5.0GiB	20.8GiB	44.5T	185.3T	
8	2.0B	2.0B	5.3GiB	21.4GiB	62.3T	259.4T	
10	2.0B	2.0B	5.6GiB	22.0GiB	80.1T	333.5T	
SD3 —Teamwork							
3	2.1B	76.3M	4.8GiB	6.8GiB	27.0T	94.2T	‡
6	2.2B	152.7M	5.5GiB	9.4GiB	54.0T	188.4T	†
8	2.2B	203.6M	5.9GiB	11.1GiB	72.0T	251.2T	★
10	2.3B	254.5M	6.4GiB	12.8GiB	90.0T	314.0T	★
SD3 —Joint Attention							
3	2.1B	76.3M	4.8GiB	6.8GiB	44.3T	172.3T	‡
6	2.2B	152.7M	5.5GiB	9.4GiB	140.7T	578.9T	†
8	2.2B	203.6M	5.9GiB	11.1GiB	233.9T	980.1T	
10	2.3B	254.5M	6.4GiB	12.8GiB	350.3T	1485.5T	
Flux —Teamwork							
3	12.1B	180.4M	24.0GiB	30.6GiB	223.9T	UNAVAIL	
6	12.3B	360.7M	25.7GiB	38.9GiB	447.8T	UNAVAIL	†
8	12.4B	481.0M	26.7GiB	44.4GiB	597.1T	UNAVAIL	
10	12.5B	601.2M	27.8GiB	49.9GiB	746.4T	UNAVAIL	
Flux —Joint Attention							
3	12.1B	180.4M	24.0GiB	30.7GiB	253.7T	UNAVAIL	
6	12.3B	360.7M	25.7GiB	38.9GiB	596.5T	UNAVAIL	
8	12.4B	481.0M	26.7GiB	44.4GiB	874.7T	UNAVAIL	
10	12.5B	601.2M	27.8GiB	49.9GiB	1192.5T	UNAVAIL	

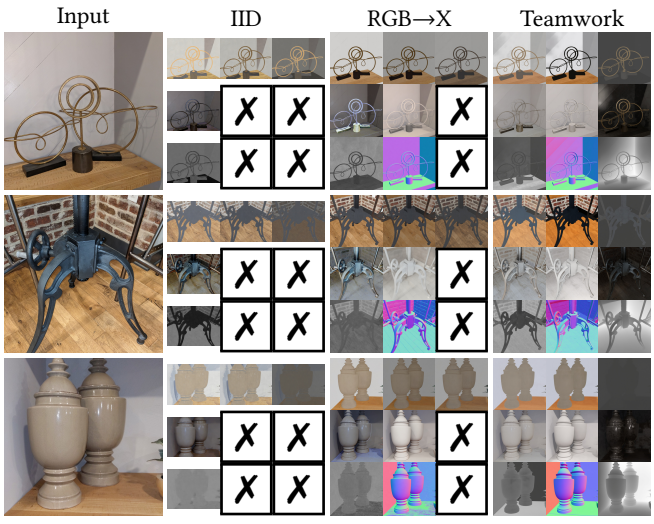


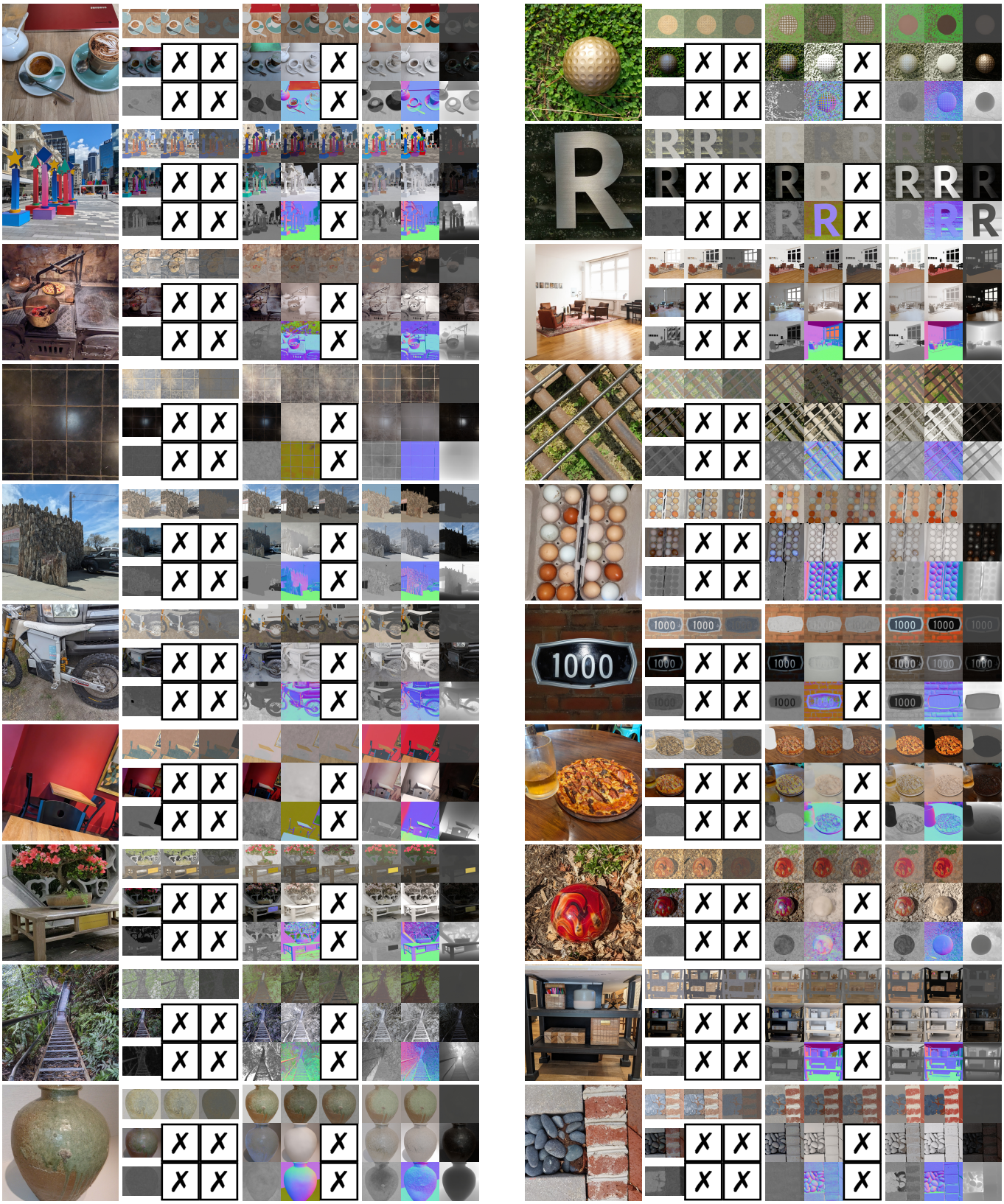


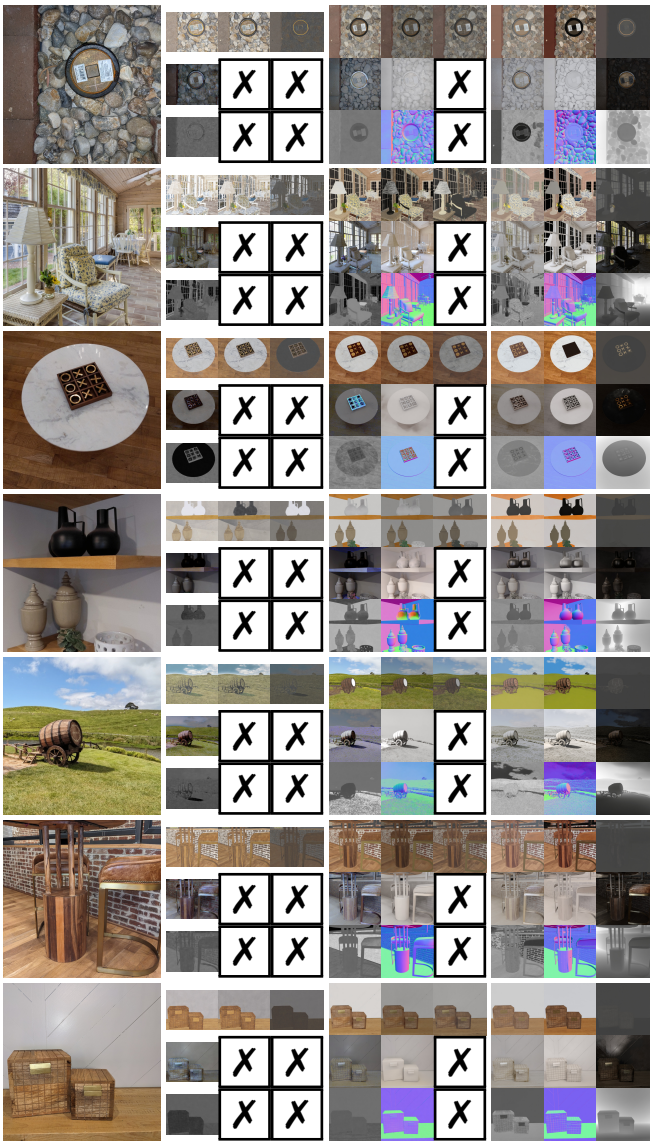


4 Real-World Results

Below are intrinsic decompositions of 30 real-world photographs. The photos show a variety of lighting conditions, shapes, and materials which are challenging for models trained only on synthetic datasets.





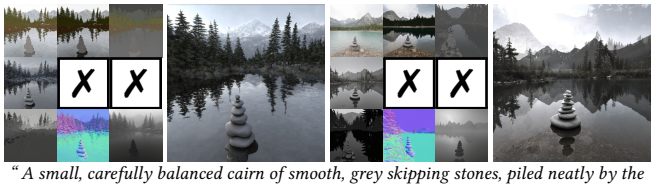


5 Intrinsic Image Synthesis Results

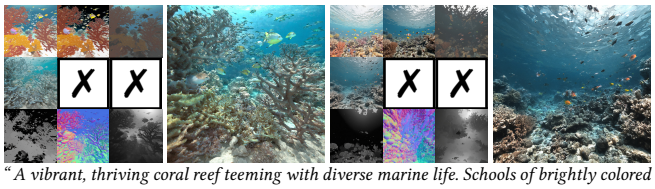
Below are 15 images generated using the intrinsic image synthesis model trained on 256k Interiorverse exemplars and prompts generated with Gemma-3. The Teamwork results are coherent and produce an image without ghosting when recomposed (i.e., (diffuse + specular) × shading). Without coordination, the maps lack coherence even when all generated from the same seed. We use same sampler and step count as in other results (Euler sampler, 50 steps) but with additional classifier-free guidance (scale 4).

"An antique, ornate silver teapot, its surface beautifully tarnished with age, resting on a rich, burgundy velvet cloth. Intricate floral patterns and swirling arabesques are etched into its bulbous form, speaking of elegant tea parties from centuries past."

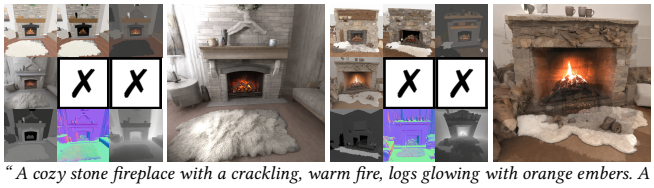
"A rustic wooden board laden with an assortment of artisanal cheeses - creamy brie, sharp cheddar, and piquant blue cheese - accompanied by bunches of juicy red grapes, sweet figs, crunchy walnuts, and slices of crusty sourdough bread, illuminated by the warm flicker of candlelight."



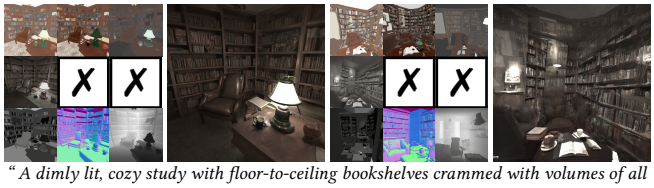
"A small, carefully balanced cairn of smooth, grey skipping stones, piled neatly by the still, reflective edge of a calm, clear mountain lake. The water mirrors the surrounding pine trees and distant, snow-capped peaks."



"A vibrant, thriving coral reef teeming with diverse marine life. Schools of brightly colored tropical fish, including clownfish and angelfish, dart playfully amongst intricate formations of brain coral, staghorn coral, and swaying sea anemones. Sunlight filters down through the clear, turquoise water, illuminating the underwater spectacle."



"A cozy stone fireplace with a crackling, warm fire, logs glowing with orange embers. A fluffy, sheepskin rug is spread invitingly before it, and a pair of empty, handcrafted pottery mugs rest on the rough-hewn wooden mantelpiece, suggesting recent, comfortable conversation."

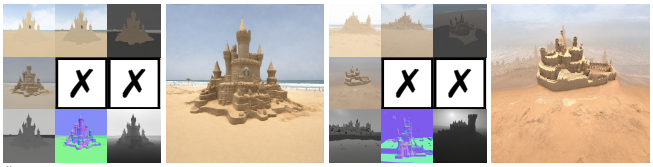


"A dimly lit, cozy study with floor-to-ceiling bookshelves crammed with volumes of all sizes. A worn leather armchair invites readers, positioned next to a grand mahogany desk where a green banker's lamp casts a warm glow on an open book and a steaming mug of tea."

“A stack of colorful, mismatched ceramic mugs on a rustic wooden shelf. One mug has a chipped rim, hinting at years of comforting morning coffees, while another boasts a quirky, hand-painted design.”

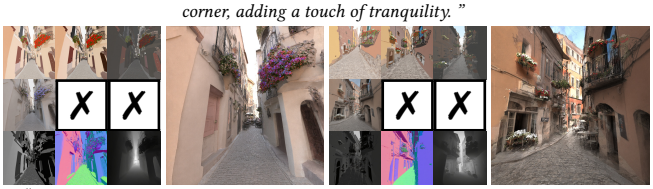
“A classic red convertible sports car, its chrome details gleaming under the bright Mediterranean summer sun, parked on a winding coastal road. The turquoise ocean stretches out to the horizon, and the wind gently ruffles the leather seats.”

“A rusty, sky-blue bicycle with a wicker basket overflowing with freshly picked wildflowers - daisies, cornflowers, and poppies - leaning against a crumbling stone wall in the sun-dappled French countryside.”

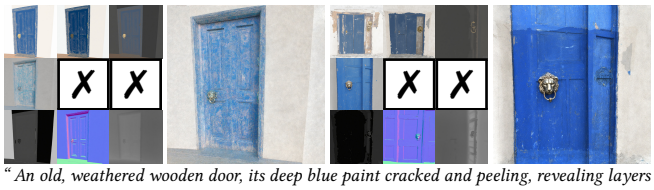


“A meticulously constructed sandcastle on a sun-drenched, golden beach, complete with towering turrets, crenellated walls, and a carefully dug moat. It stands proudly, just moments before the incoming tide begins to reclaim it.”

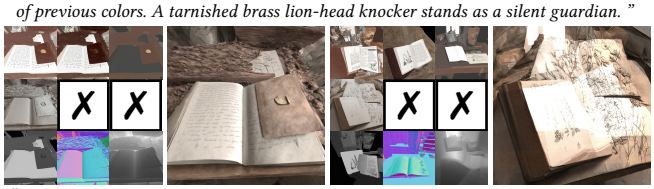
“A serene yoga studio, bathed in the soft, diffused light of early morning filtering through large, frosted windows. Pale wooden floors are adorned with neatly rolled yoga mats in muted earth tones. A single, elegant white orchid sits on a low, dark wood table in the corner, adding a touch of tranquility.”



“A charming, narrow cobblestone street in an old Tuscan town, lined with colorful, sun-bleached buildings. Flower boxes overflowing with geraniums adorn wrought-iron balconies, and quaint cafes with outdoor seating invite passersby.”



“An old, weathered wooden door, its deep blue paint cracked and peeling, revealing layers of previous colors. A tarnished brass lion-head knocker stands as a silent guardian.”

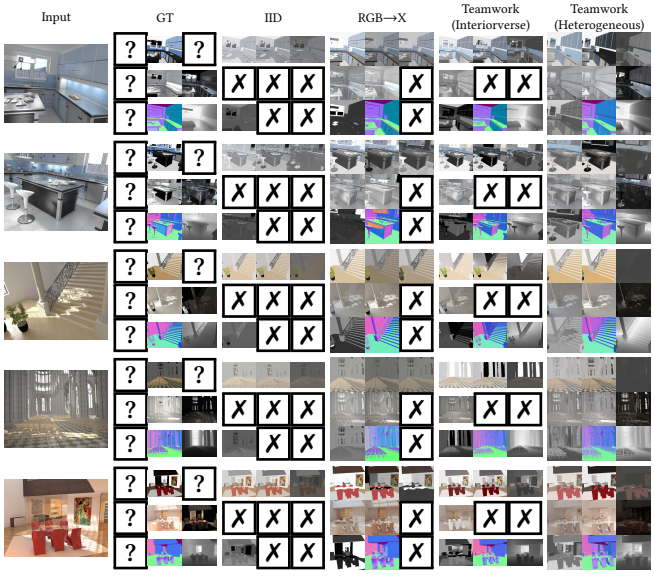


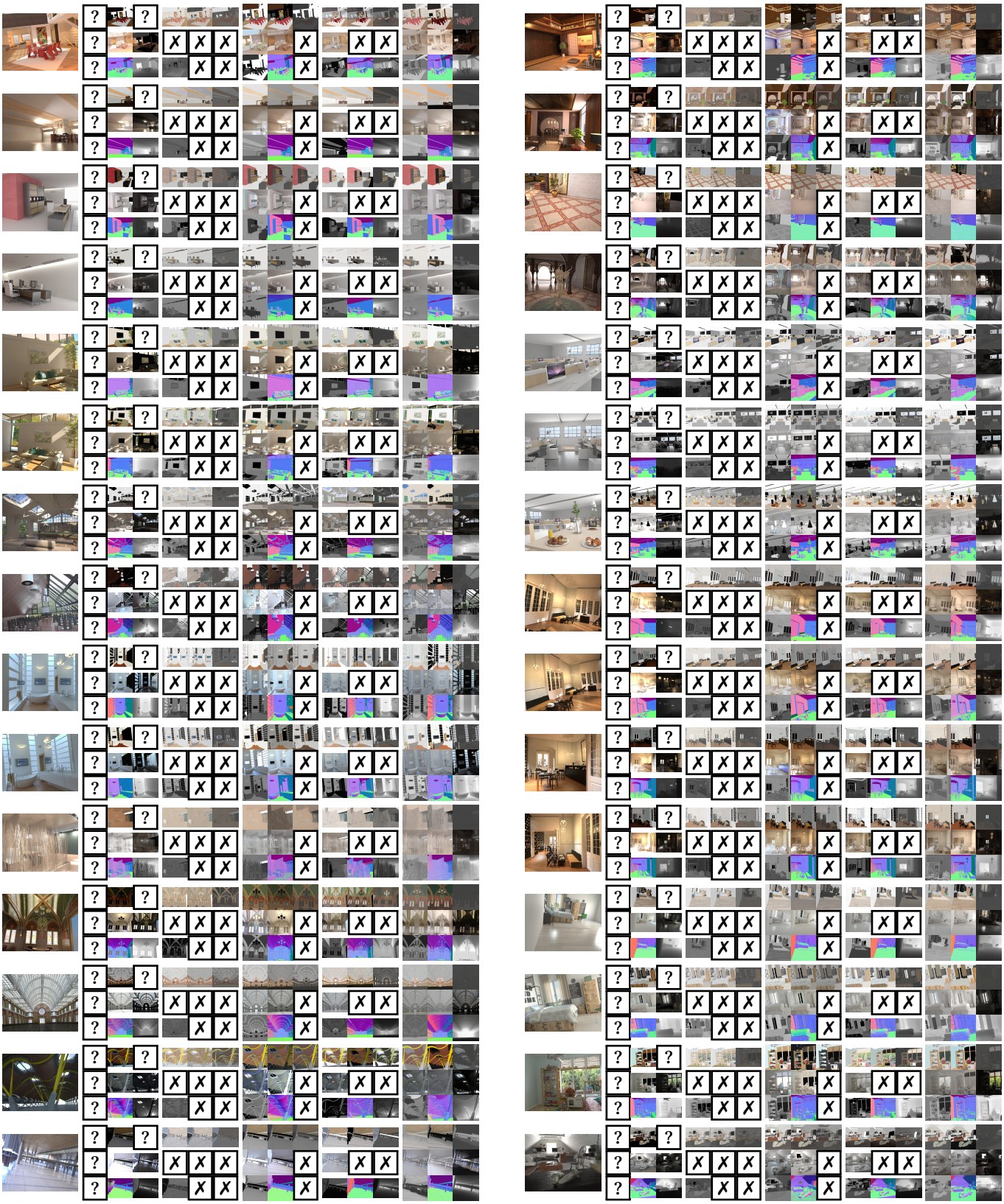
“A worn, leather-bound journal, its cover softened and creased with age and use. It lies open on a rustic, ink-stained wooden desk, its thick, cream-colored pages filled with elegant, flowing cursive script and faded ink sketches of botanical specimens.”

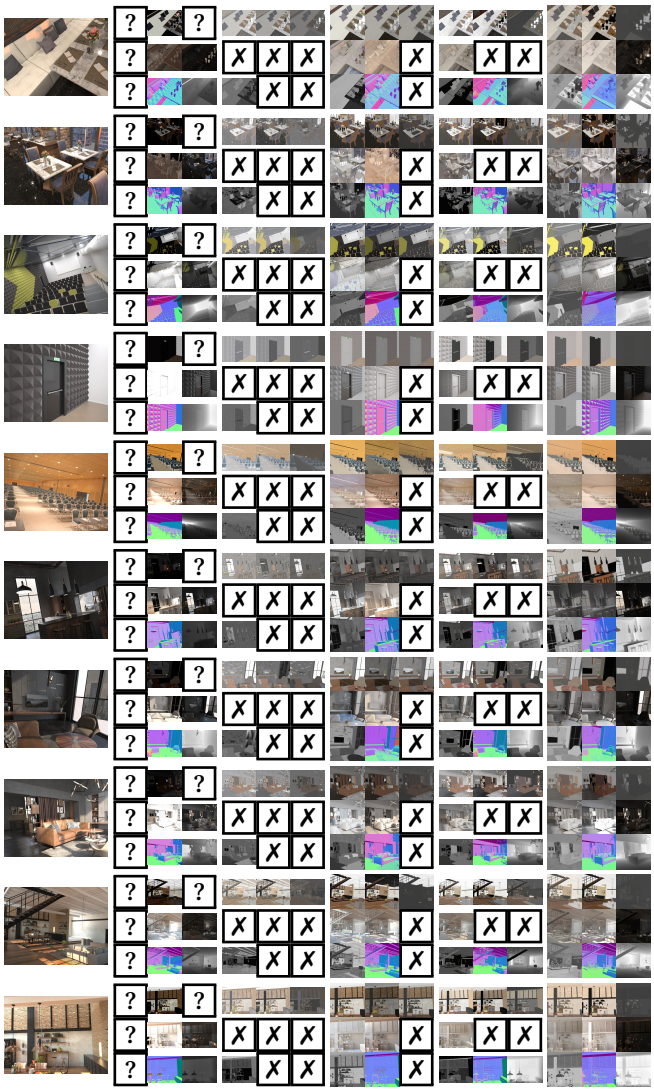
“A worn leather satchel, its brass buckles gleaming with a soft patina, resting on a cobblestone street in an old European city. A corner of a faded, hand-drawn map peeks out from an unfastened flap, promising adventure.”

6 HyperSim Results

Below are results for a subset of 60 frames from the HyperSim test set (roughly one per scene). Models trained exclusively on InteriorVerse all use a rescaling to 640 × 480 (InteriorVerse's resolution) from the HyperSim's native 1024 × 720. The published RGB→X model does not operate well at that resolution, and so is evaluated on (and against) a 1024 × 1024 scale+crop. We use the same resolution for our heterogeneous model, as it performs comparably at a range of resolutions.







7 InteriorVerse Results

Below are results for a subset of 60 frames from the InteriorVerse test set (roughly one per scene). Models trained exclusively on InteriorVerse all use the dataset's native resolution of 640×480. The published RGB→X model does not operate well at that resolution, and so is evaluated on (and against) a 1024×1024 scale+crop. Our heterogeneous model performs comparably at a range of resolutions, so we display results at the native 720×480.