Generating Visual Scenes from Touch
– Supplementary Material –

We provide additional details about our method, and provide qualitative results for our generation tasks.

A. Model Architecture and Implementation Details

We provide additional details about the latent diffusion model, such as the training hyperparameters.

Table 1: We show detailed hyperparamters setting of our models, including first stage model, condition model and LDM model.

Hyperparamter	Value	Hyperparamter	Value
Learning Rate	2×10^{-6}	LDM Model	U-Net
Image Size	256	LDM Input Size	64
Channel	3	LDM Input Channel	3
Conditioning Key	Crossattn	LDM Output Channel	3
First Stage Model	VQModelInterface	LDM Attention Resolutions	[8,4,2]
VQ In-channel	3	LDM Num Resblocks	2
VQ Out-channel	3	LDM Channel Mult	[1,2,3,5]
VQ Num. Resblocks	2	LDM Num Head Channels	32
VQ dropout	0.0	LDM Use Spatial Transformer	True
Condition Model	CVTP ResNet-18	LDM Transformer Depth	1
Condition Layer	5	LDM Context Dim	512
Condition Frame	5	Batch Size	48
Cond Stage Trainable	True	Monitor	val/loss_simple_ema
Diffusion Timesteps	1000	Epoch	30
Scheduler	DDPM		

B. More Qualitative Results

We provide additional results visuo-tactile cross generation, tactile-driven stylization and tactile-driven shading estimation.

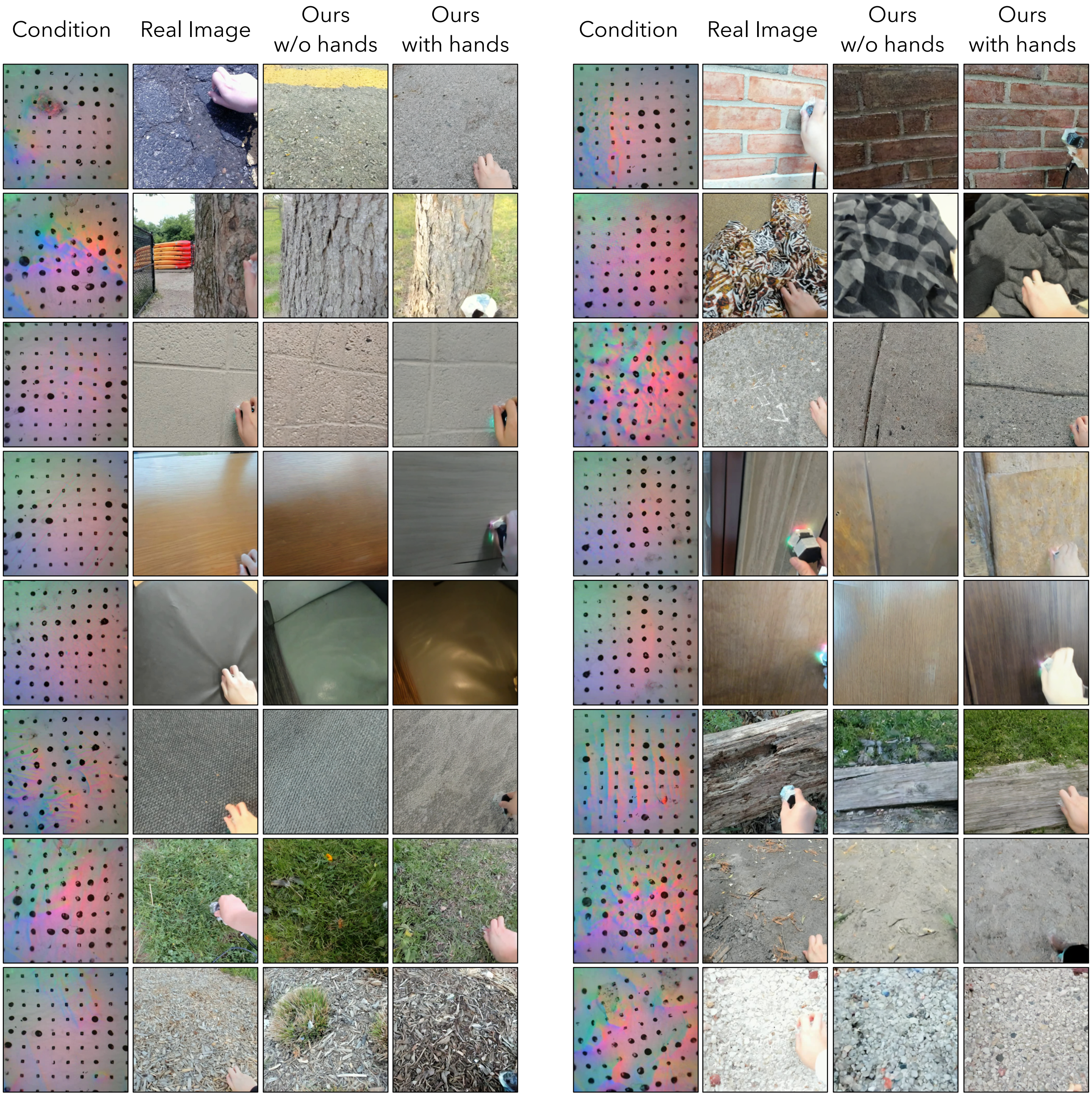

Figure 1: Additional results for **touch-to-image generation** on *Touch and Go* dataset, where we show both our results with and without sensors.

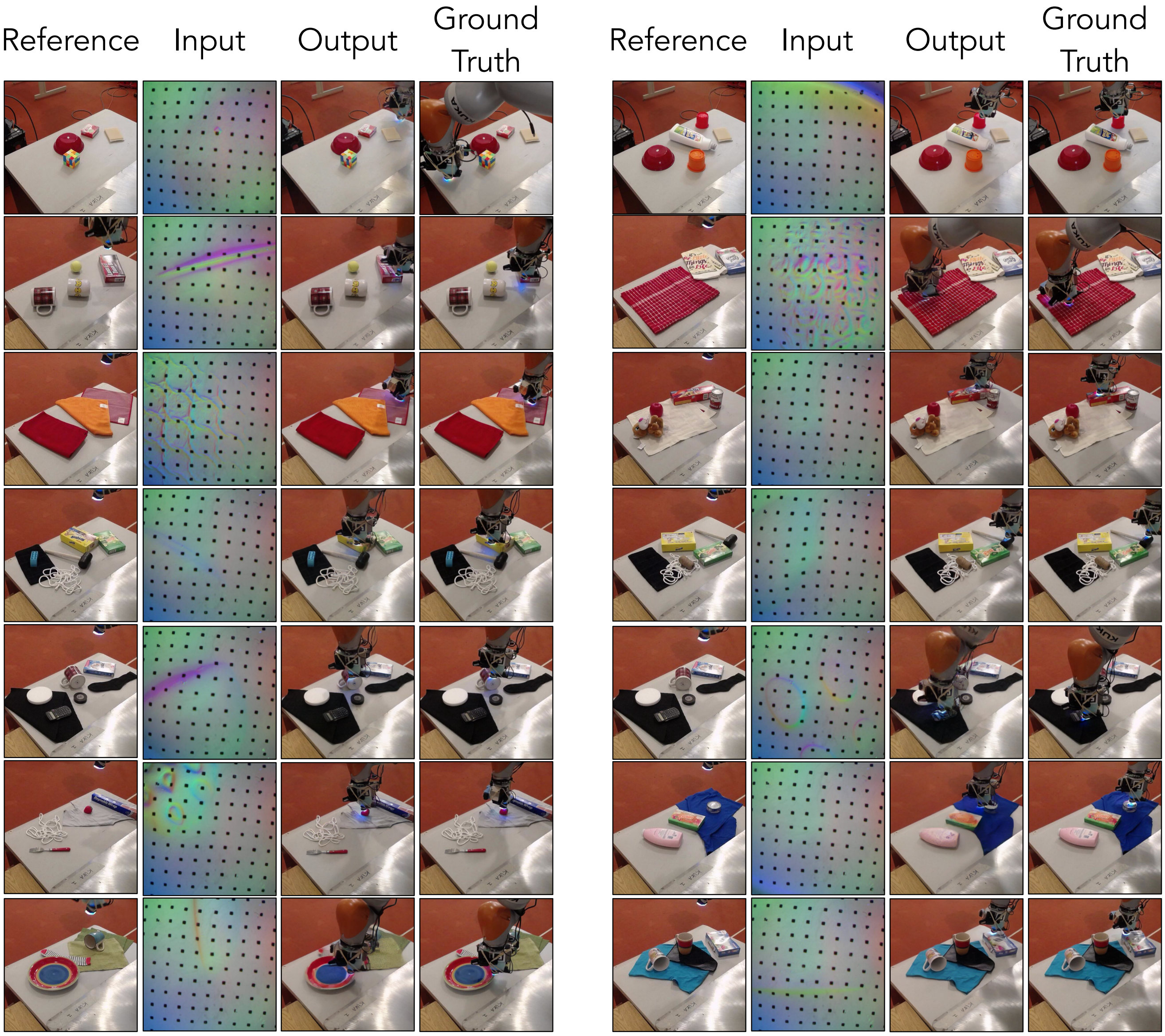

Figure 2: Additional results for **touch-to-image generation** on *VisGel* dataset.

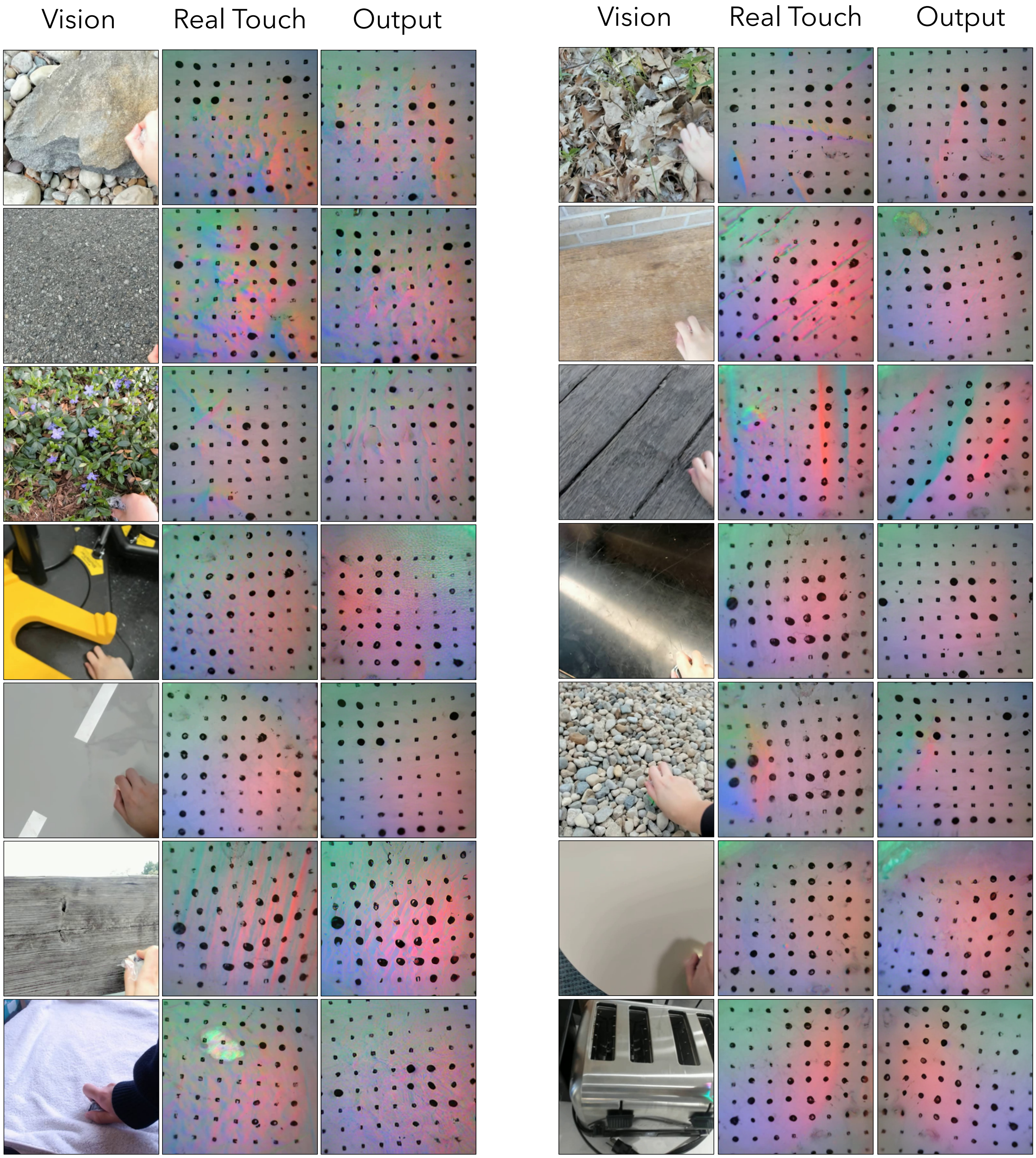

Figure 3: Additional results for **image-to-touch generation** on *Touch and Go* dataset.

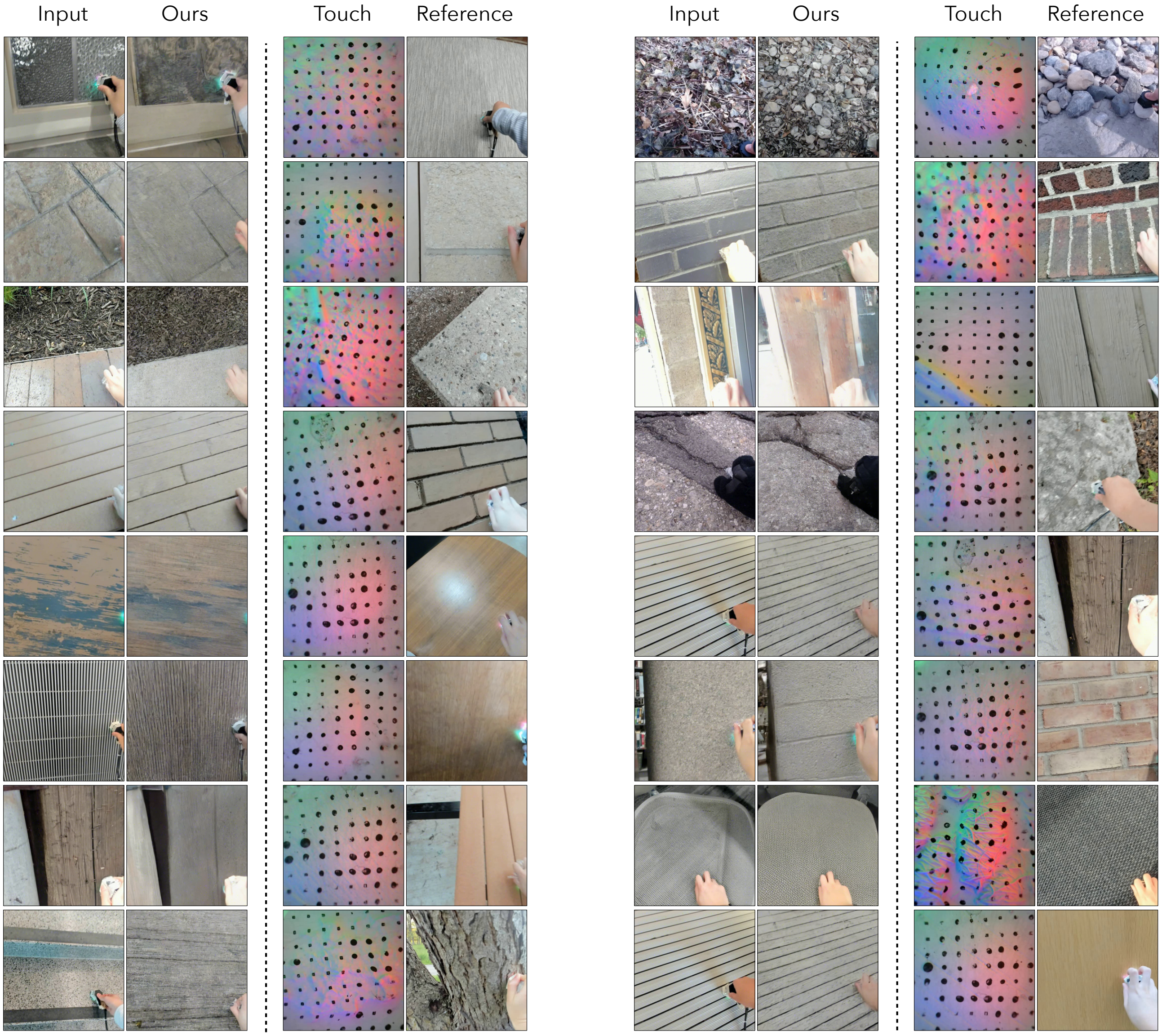

Figure 4: Additional results for **tactile-driven image stylization** results. (*Zoom in for better viewing*)

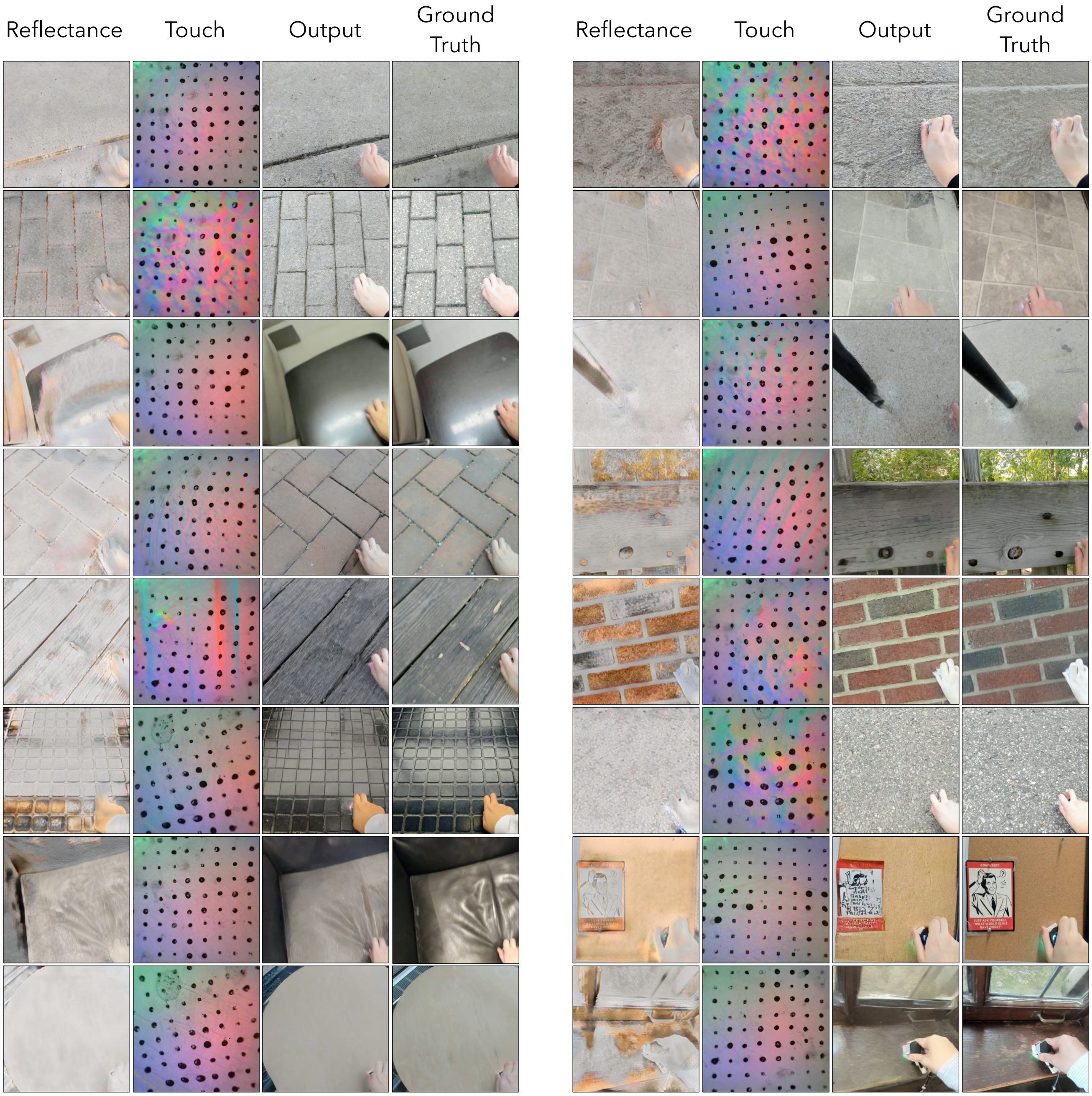

Figure 5: Additional results for **tactile-driven shading estimation**. (*Zoom in for better viewing*)