

Supplementary Materials of Scene Diffusion: Text-driven Scene Image Synthesis Conditioning on a Single 3D Model

Anonymous Authors

1 FURTHER DISCUSSION ON THE SHADING ADAPTIVE TRANSFORMATION \mathcal{T}

In the section of methodology, we outline how the shading adaptive transformation \mathcal{T} functions within the SACA to reconcile the contradiction between two training objectives and facilitate network’s learning on the prior of global shading coherence. In support of this argument, we offer an intuitive comparison in Figure 1. The training hyperparameters are the same for both setups, saving for the exclusion of \mathcal{T} in the first one. As can be seen, not utilizing \mathcal{T} results in the decreased shading diversity for the object across the various scenes. The object may seem dark even in the scene with bright lighting. As a contrast, the object’s shading exhibits more dynamic changes when \mathcal{T} is employed.

We also include a brief analysis in the main text to confirm that the utilized \mathcal{T} can effectively handle the variations in the ambient light color of the object. Figure 2 displays visual illustrations to support this assertion. Take case 1-3 as the examples, when object is in the scene with relatively uniform lighting, \mathcal{T} can largely eliminate the shading difference between c and x_0 . However, since the current \mathcal{T} does not take diffuse reflection and specular highlight colors into account, the shading variances they initiated are still hard to eliminate, especially in the samples as case 4. This is also a topic that we plan to explore in the future research. Overall, the current form of \mathcal{T} is concise and theoretically rational. The application of it has been shown to yield beneficial results for the network’s learning.

2 VISUAL DEMONSTRATION OF THE FPTs’S EFFECT

In the main text, the EoG indicator is employed to quantitatively verify that FPTs can efficiently reduce the high-frequency signals in the object area. Figure 3 further provides the visual evidences to uphold this assertion. The top and bottom rows exhibit the sequences of predicted \hat{x}_0 during the denoising process, without or with FPTs, respectively. In the absence of FPTs, the high-frequency components in the object area will grow quickly. Though there is a decline in the subsequently, they still be unreasonably high in the end. The introduction of FPTs alleviates this problem. As depicted in the figure, it softens the growth of high frequency components in object area, regulating them to the reasonable extent eventually.

3 FURTHER INTRODUCTION TO THE DATASET

The dataset used in this work is constructed based on 3D-FUTURE. It consists of 20240 high-quality interior design rendering images and the textured 3D models of included furniture. Condition images are create based on Blender 3.6. We setup the object and camera positions according to the annotation, place a daylight source above them, and then execute the single-model rendering. Since a scene

image usually contains multiple furniture, it will correspond to multiple condition images. The text prompt corresponding to each scene image is obtained based on the annotation and LLaVA-1.5. After leaving out the test objects and the scene images containing them, there are 19127 different scene images and 49,963 condition-text-output triples involved in training. The condition images can be divided into five categories according to the furniture they contain: bed (4148), sofa (12259), table (4129), chair (5195) and shelf (24232). Figure 4 shows some examples of the training data.

4 IMPLEMENTATION DETAILS OF THE COMPARISONS

Four alternative methods are selected for comparison in this paper. This part provides the implementation details of them. All these methods are diffusion-based. Unless otherwise specified, all methods use the same sampling setup as the proposed method, where 100-step DDIM sampler and the classifier-free guidance scale of 7 are used.

When performing **BLIP-Diffusion**, to achieve the control over the pose of objects, the recommended paradigm¹ that integrating with ControlNet is used. Concretely, we use the original condition image as the input to subject encoder, and the canny map of it as the input to ControlNet.

For **SD-Inpainting**, we use the implementation provided by Stable Diffusion WebUI². Masked content is set to *latent nothing*, inpaint area is set to *whole picture*, and denoising strength is 1.

For **InstructPix2Pix**, we use the online demo³ provided by the original authors. Text CFG and image CFG are set to 7.5 and 1.5 according to the recommendation. The editing instruction is given as “Change the background to...”.

For **ControlNet**, we use our dataset to train it based on its original codebase⁴. The base network is Stable Diffusion V2.1. Except for the utilization of \mathcal{L}_{saca} , all the training hyperparameters are same as the proposed method.

5 MORE VISUAL RESULTS

This part provides more visual illustrations about the experiments. Figure 5,6 show more results about the application effect of scene diffusion. Figure 7 shows more results about the comparison with existing alternatives. Figure 8 shows the original images of the results exhibited in the ablation study of main text. Figure 9,10 show more results about the expanded applications of the proposed method.

¹https://huggingface.co/docs/diffusers/main/en/api/pipelines/blip_diffusion

²<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

³<https://huggingface.co/spaces/timbrooks/instruct-pix2pix>

⁴<https://github.com/llyasviel/ControlNet>

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

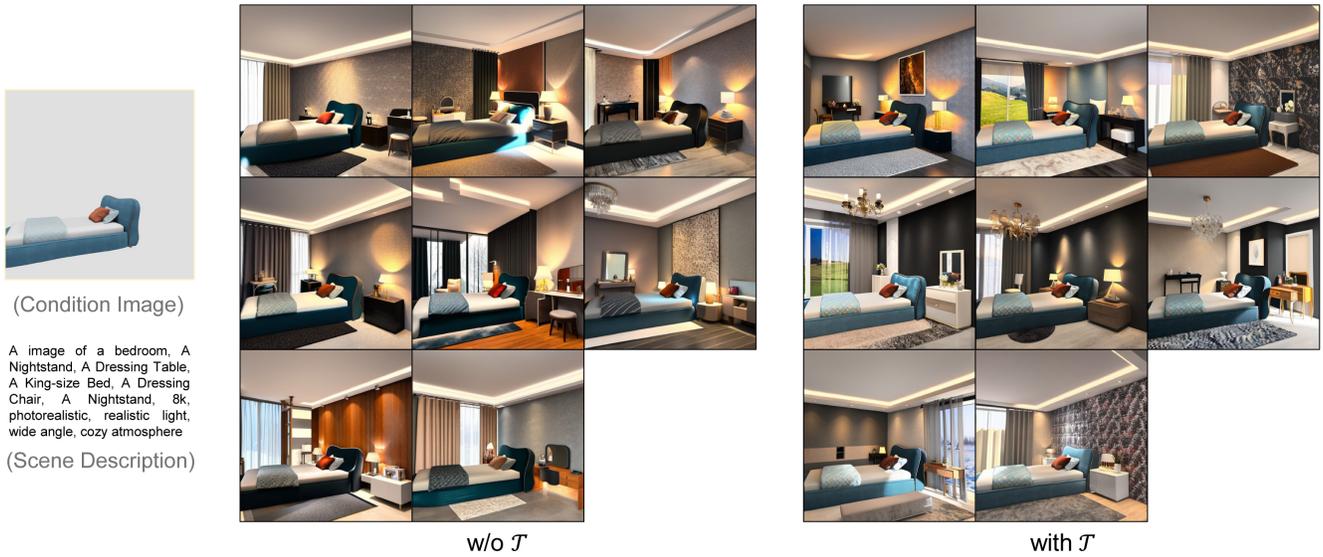


Figure 1: The comparison between the setups with or without the shading adaptive transformation \mathcal{T} . Each group of images are produced in the same mini-batch.

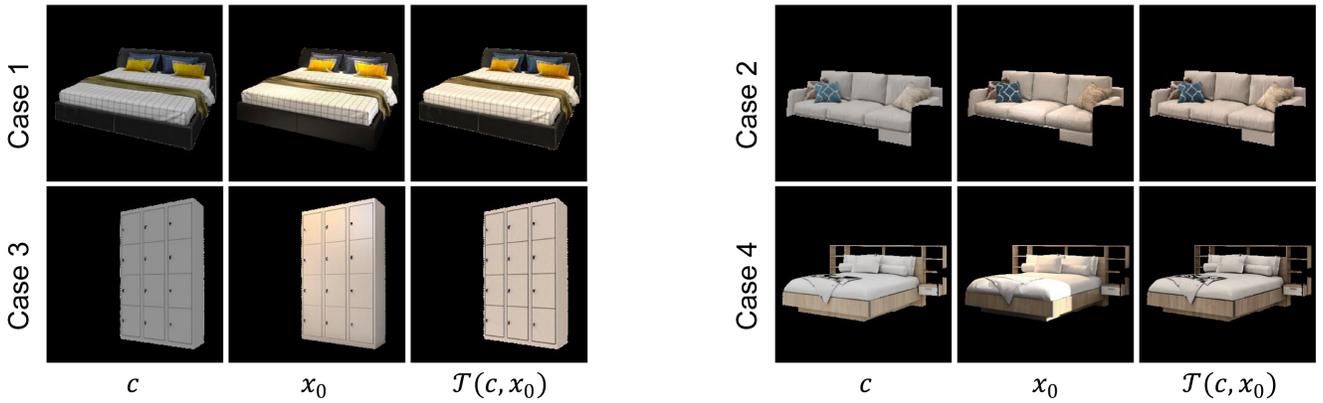


Figure 2: The effect of shading adaptive transformation \mathcal{T} . Since \mathcal{T} is solely conducted on the object areas, the background areas in these images are omitted.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348



Figure 3: The comparison between the sequences of predicted \hat{x}_0 with or without FPTS.

Scene image	Condition Images		Scene image	Condition Images	
					
Scene Description A living room with a coffee table in the center, surrounded by a yellow multi-seat sofa, an office chair, a wooden sideboard, and a wooden sofastool, with a modern pendant lamp hanging above, creating a cozy and inviting atmosphere.			Scene Description A bedroom featuring a wardrobe, a black king-size bed, a round nightstand, a rococo dressing table, and a dressing chair, with a pendant lamp above, adding a tranquil ambiance to the room.		

Figure 4: The examples of constructed condition-text-output data.

Single-object



A detailed depiction of a bedroom scene, showcasing a single bed, a desk, and an office chair, each positioned to complement the room's overall ambiance.



A living room composition, showcasing a loveseat sofa and an armchair, arranged to create a harmonious and inviting space.



A living room scene, featuring dining chairs and an accompanying dining table, arranged to evoke a welcoming atmosphere.



A living room visualized with a table, a corner table, a loveseat sofa, and a sofastool, each element thoughtfully positioned to craft a cohesive and inviting atmosphere.



A living room scene, elegantly set with a dining table, dining chairs, and a sideboard, curated to blend seamless functionality with sophisticated style.



A depiction of a living room furnished with an office chair, a TV stand, a nightstand, and a corner table, arranged to balance functionality with aesthetic appeal.

Multi-object



A living room setting with a king-size bed as the focal point, accompanied by a TV stand and a nightstand.



A living room setting, highlighted by an armchair, a corner cabinet and a corner table, meticulously placed to enhance the space's comfort and aesthetics.



A living room featuring a multi-seat sofa and an armchair, arranged for cozy gatherings.



A bedroom with a king-size bed, wardrobe, office chair, and white walls, arranged for elegance and comfort.

Figure 5: More results about the application effect of Scene Diffusion in Single-object and Multi-object scenarios.

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522

523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580

Variable Scene Description Text



(Condition Image)

- ◆ Object Ctrl.
- ◆ Style Ctrl.
- ◆ Color Ctrl.
- ◆ Material Ctrl.



A bright, airy living room with a sofa, large windows and minimalist black and white artwork on the walls.



A cozy setup with soft lighting, a sofa, a sleek glass coffee table, plush area rugs, an art paintings of human faces.



A modern industrial living room with a sofa, high ceilings, black shelving, and exposed brick walls with metal accents.



The cozy living room with soft lighting, a sofa, thick plush area rugs, introducing open shelving.



A Scandinavian-inspired living room with natural light, a sofa, a sofastool and a white rugs.



A modern simple living room with a sofa and an artwork on the gray wall, introduces indoor greenery for a fresh touch.



A colorful, boho-chic living room with a sofa, vibrant rugs, wooden flooring, a small table.



The minimalist living room, with a sofa and concrete floors, adds a black shelving unit displaying modern art pieces.



A warm, earth-toned living room with a sofa, wooden flooring, and the geometric shelving filled with ceramic and wooden decor.



A luxurious living room with a sofa, plush textiles, the modular bookcase, and accents in deep blues and gold.

Variable Position & Posture of Object

A living room detailed with a sofa, soft white walls, gray ceramic tile floor, a coffee table, a office chair, harmonious layout of the whole room.

(Scene Description)



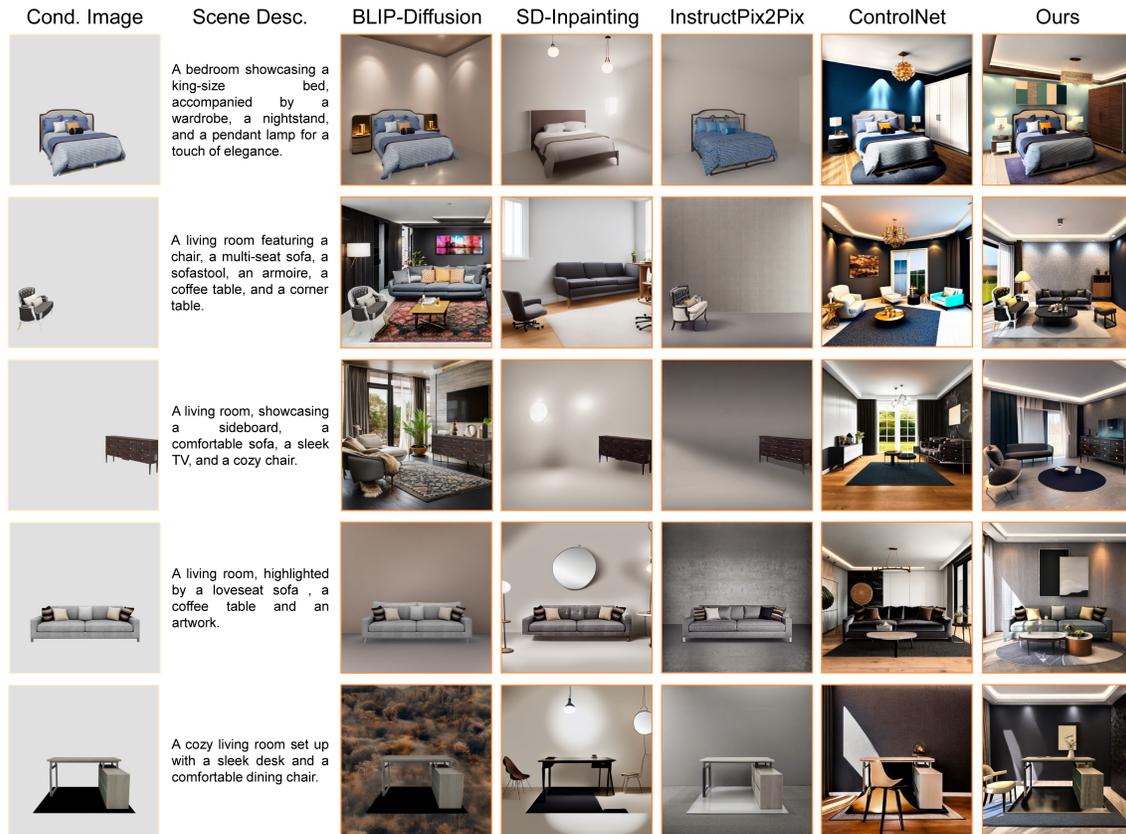
A bedroom with a king-size bed, showcasing soft pink walls, white ceramic floor, a nightstand and a sideboard.

(Scene Description)



Figure 6: More results about the application effect of Scene Diffusion in Variable Scene Description Text and Variable Position & Posture of Object scenarios.

581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638



639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696

Figure 7: More results about the comparison with the existing alternatives.

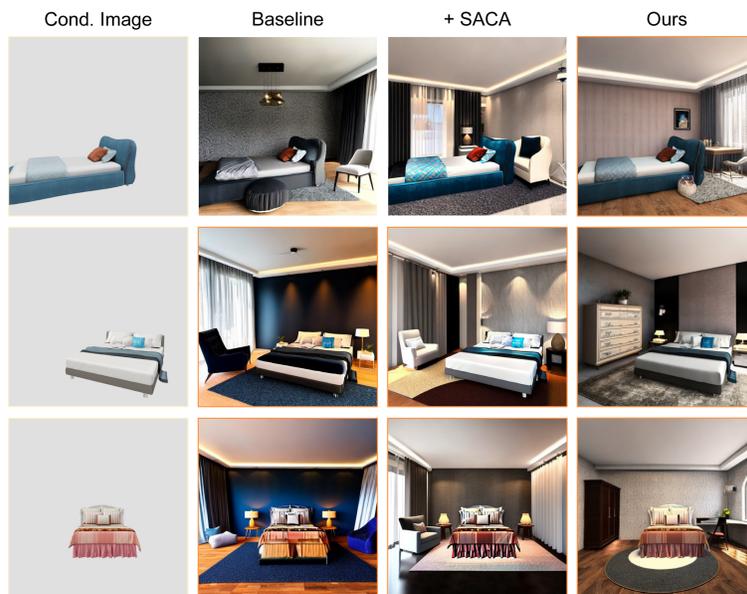
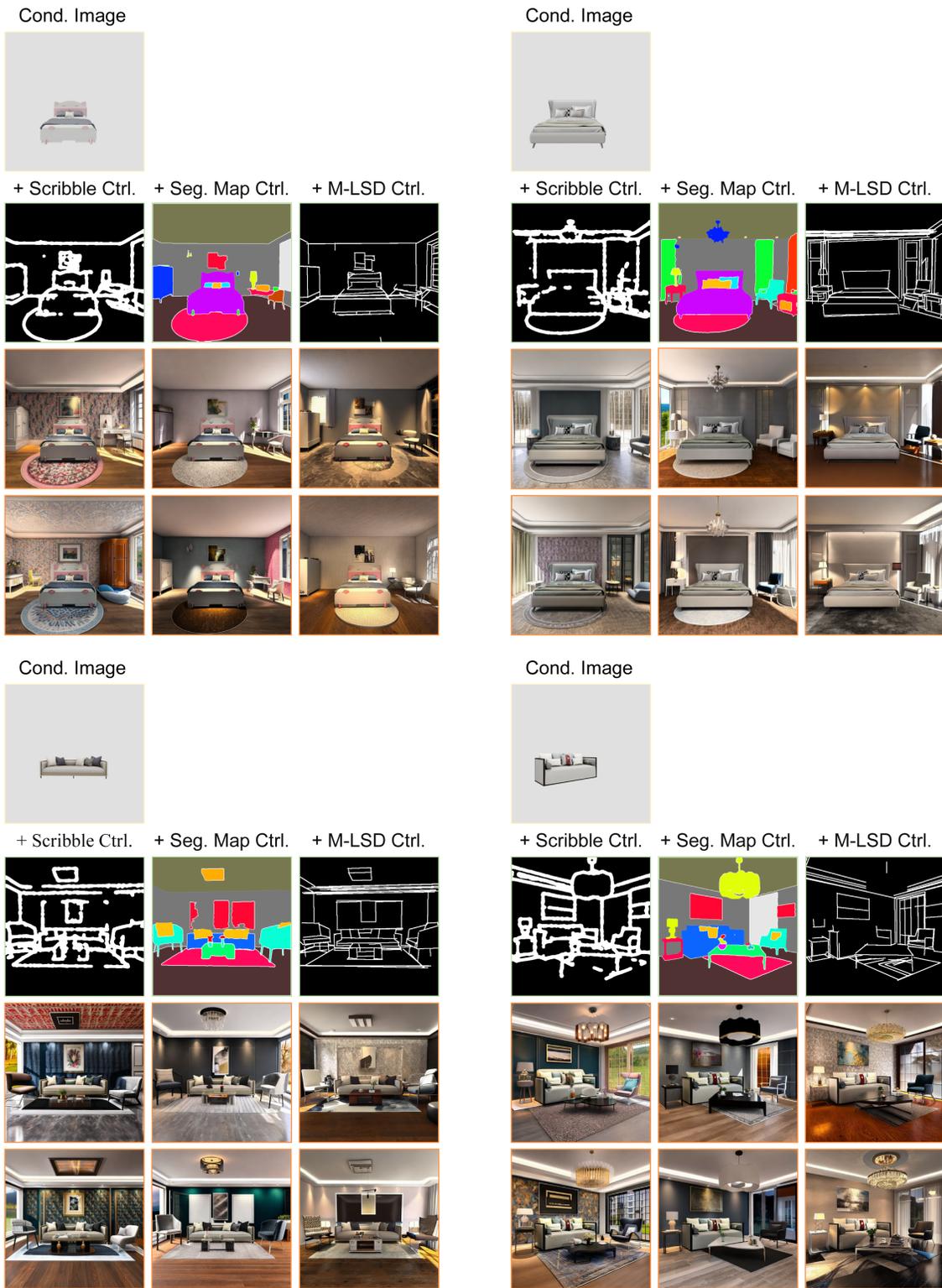


Figure 8: The original images of the results exhibited in the ablation study of main text.

697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754



755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812

Figure 9: More results about the expanded application of Scene Diffusion in Integrating with Existing ControlNet.

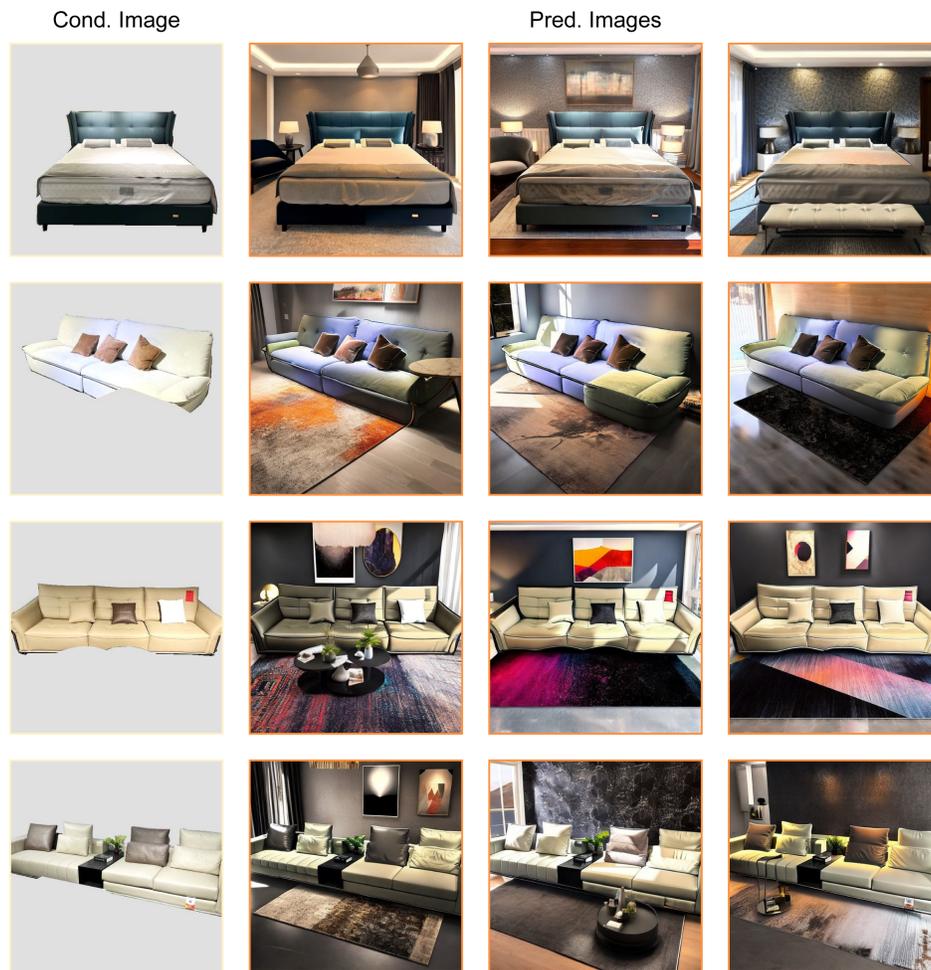


Figure 10: More results about the expanded application of Scene Diffusion in Generalizing to Real Image Fragment.