

SUPPLEMENTARY MATERIAL FOR PAPER: R-GENIE: REASONING-GUIDED GENERATIVE IMAGE EDITING

Anonymous authors

Paper under double-blind review

To enhance the reproducibility and transparency of our work, we present additional dataset samples in Section A1. We also provide comprehensive visualizations of the ablation study results in Section A2, along with a comparative analysis against contemporary unified multimodal understanding and generation approaches in Section A3. After that, more experiments on runtime evaluation are provided in Section A4, meanwhile more convincing qualitative comparison results are illustrated in Section A5. Furthermore, we provide details about user study results in Section A6.

A1 DATASET DETAILS

As detailed in Section 3.2, our dataset features curated instruction-image-edit triples for image editing tasks, where each sample incorporates natural language instructions requiring compositional reasoning, source and target image pairs. Here we present more triples examples of REditBench in Figure A1. Besides, the comparison with related datasets is presented in the Table A1. It is also worth mentioning that our dataset is released under the CC BY-NC 4.0 (research-only) with commercial use requiring authorization.

Limitations in domain coverage. While our approach generalizes well to in-distribution edits (*e.g.*, modifying common objects like “persons” or “dogs”), performance may degrade for highly specialized domains (*e.g.*, medical imaging or rare artistic styles) where training data is scarce. Future work could incorporate domain adaptation methods to mitigate this limitation.

The potential bias. Potential biases stem from template structure (prioritizing certain reasoning patterns) and annotator tendencies (*e.g.*, frequent attribute combinations like “red car”). We mitigated this via diverse annotators but acknowledge some cultural/linguistic biases may persist.

Table A1: Comparison of different datasets.

Dataset	Controllable	Reasoning	Size	Open Access
InstructPix2Pix (Brooks et al., 2023)	✓	✗	454,445	✓
Reason50K (He et al., 2025)	✓	✓	51,039	✗
ReasonPix2Pix (Jin et al., 2024)	✓	✓	40,212	✗
MagicBrush (Zhang et al., 2023)	✓	✗	10,388	✓
EditWorld (Yang et al., 2024)	✓	✗	10,000+	✓
RISEBench (Zhao et al., 2025)	✓	✓	360	✓
ReasonEdit (Huang et al., 2024)	✓	✓	219	✓
REditBench (Ours)	✓	✓	1,070	✓

A2 VISUALIZATION OF ABLATION STUDY RESULTS

To rigorously evaluate the performance gains attributable to various components of our proposed paradigm, Figure A2 presents a comprehensive ablation study through comparative visual analysis. The systematic integration of individual architectural elements (from left to right) demonstrates statistically significant improvements in the model’s cross-modal reasoning capabilities. Quantitative metrics confirm that each progressive enhancement: (1) elevates semantic alignment accuracy between

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107









	
<p>"In the picture, which one is a input device for computer? Let it disappear."</p>	<p>"Which is the guy's reflection? Turn his reflection into Taylor Swift."</p>
	
<p>"There is a business meeting to be held. Whose outfit looks more formal? Exchange that person with Bill Gates."</p>	<p>"In this picture, who cannot slide on the ground? Turn him into a pole."</p>
	
<p>"If you want to call someone, which one will you use? Take it away from the picture."</p>	<p>"Who is dribbling the ball in the game? Take him away from the picture."</p>
	
<p>"What object help the man slide on the ground? Let the man wear a pair of roller skates."</p>	<p>"Who have the same exterior color as a panda? Turn them into giraffes."</p>

Figure A1: More Examples of the annotated image-instruction-edit triples.

108
109
110
111
112
113
114
115
116
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

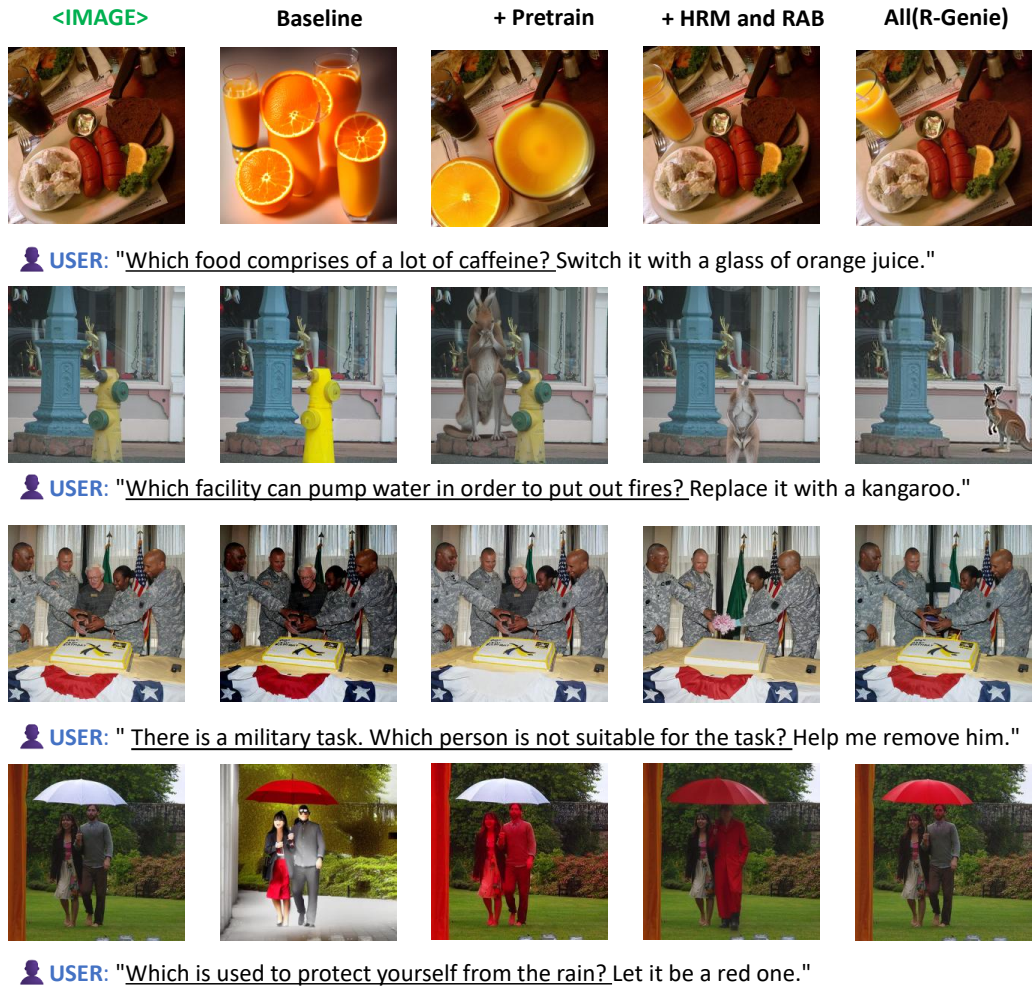
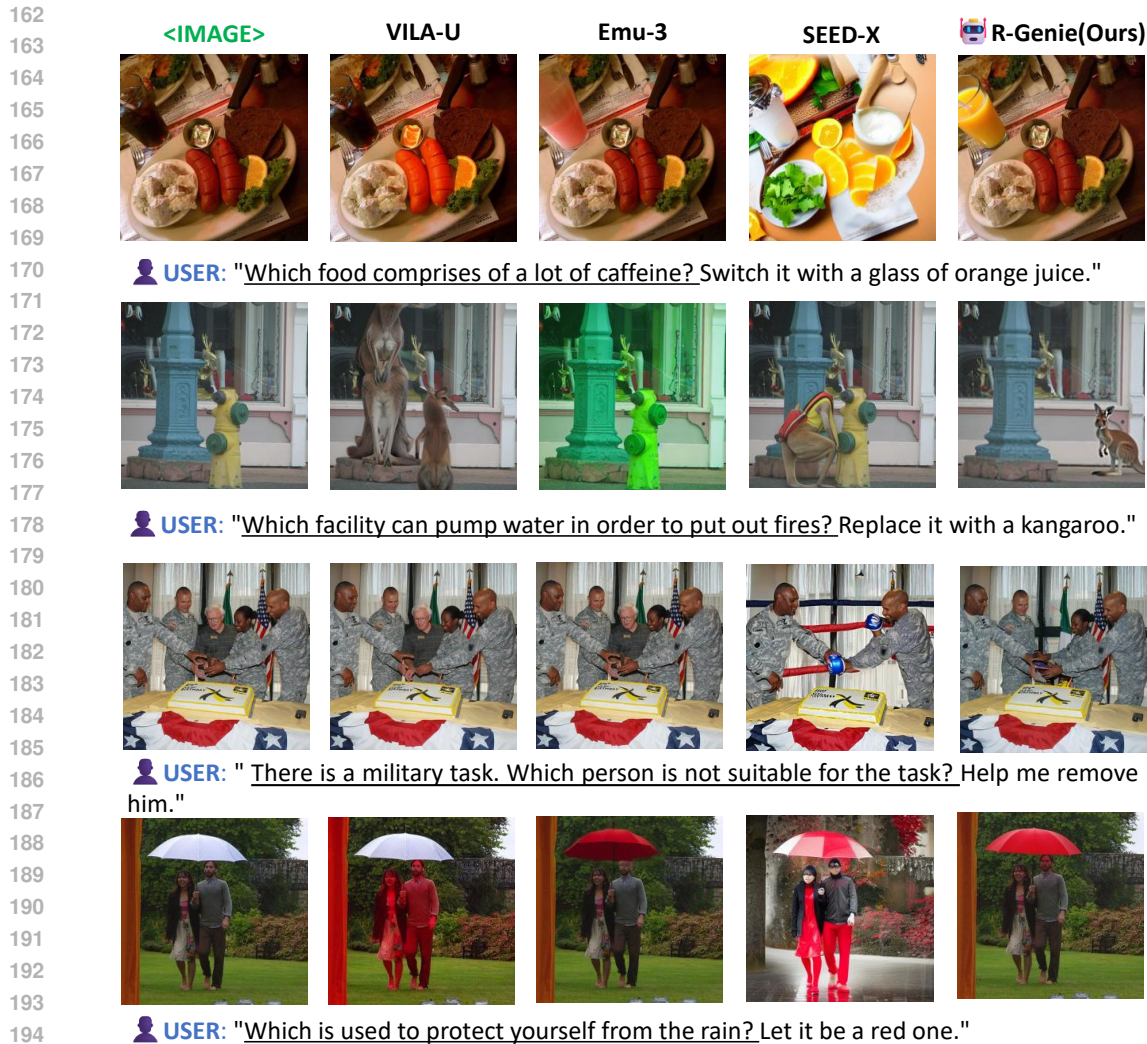


Figure A2: Visualization of Ablation Study Results.



196 Figure A3: Qualitative comparison with unified multimodal understanding and generation methods.

197

198

199 input modalities, and (2) enhances perceptual coherence in synthetic outputs. These empirical results

200 validate our design choices while providing insights into the relative contributions of each module.

201

202

203 **A3 COMPARATIVE ANALYSIS WITH UNIFIED MULTIMODAL UNDERSTANDING**

204 **AND GENERATION METHODS**

205

206

207 To ensure comprehensive comparative analysis, we extend our evaluation to benchmark general

208 multimodal understanding and generation frameworks (Wu et al., 2024; Xiao et al., 2024; Wang et al.,

209 2024; Ge et al., 2024), despite their inherent limitations in being specifically optimized for instruction-

210 guided image editing tasks. As demonstrated in Figure A3, VILA-U (Wu et al., 2024) exhibits

211 fundamental deficiencies in target object recognition across most test samples, indicating critical

212 limitations in visual grounding capabilities. However, Emu-3 (Wang et al., 2024) produces outputs

213 with marginal modifications relative to source images, revealing constrained multimodal reasoning

214 and adaptive generation capacities. SEED-X has shown its identification capabilities in certain

215 scenarios and is still limited in background perseverance. These comparative observations collectively

illustrate the technical challenges in achieving robust integration of semantic understanding and

precise image manipulation within current multimodal frameworks.

Table A2: Comparison of different datasets.

Methods	Runtime (generating one sample)	Param
InstructPix2Pix (Brooks et al., 2023)	26.6s	4.1B
OmniGen (Xiao et al., 2024)	24.6s	3.8B
SmartEdit (Huang et al., 2024)	43.5s	7.0B
RGenie(Ours)	14.7s	1.3B



Figure A4: More visual comparison results.

A4 MORE EXPERIMENTS

In this section, we compared the runtime required for the model to generate one sample. As shown in R-Table A2, our obtained results (evaluated by two NVIDIA GeForce RTX 3090 GPUs) reveal that R-Genie’s runtime is highly competitive: requiring only 14.7s per sample under the use of Phi-1.5 (Li et al., 2023). This is significantly faster than comparable methods like InstructPix2Pix (26.6s), OmniGen (24.6s), and SmartEdit (43.5s), while delivering substantial accuracy gains as demonstrated in Table 1 of the main paper.

A5 MORE QUALITATIVE COMPARISON RESULTS

To present more convincing and comprehensive qualitative comparisons, here we provide more editing samples in Figure A4.

A6 USER STUDY RESULTS

To evaluate the efficacy of our methodology, we conducted a comprehensive user study as detailed in the Experiment section. Each response alternative in the survey corresponds to one of the compared methods: Option A denotes outputs from InstructPix2Pix (Brooks et al., 2023), Option B represents results generated by MGIE (Fu et al., 2023), and Option C indicates outcomes from R-Genie. The aggregated results presented in Figures A5 to A10 demonstrate that R-Genie achieved statistically significant preference among participants. This empirical validation substantially supports our method’s superiority in human perceptual evaluation compared to existing baseline approaches.

REFERENCES

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*, pp. 18392–18402, 2023. 1, 5

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



Figure A5: User Study Snapshot: Page 1

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
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

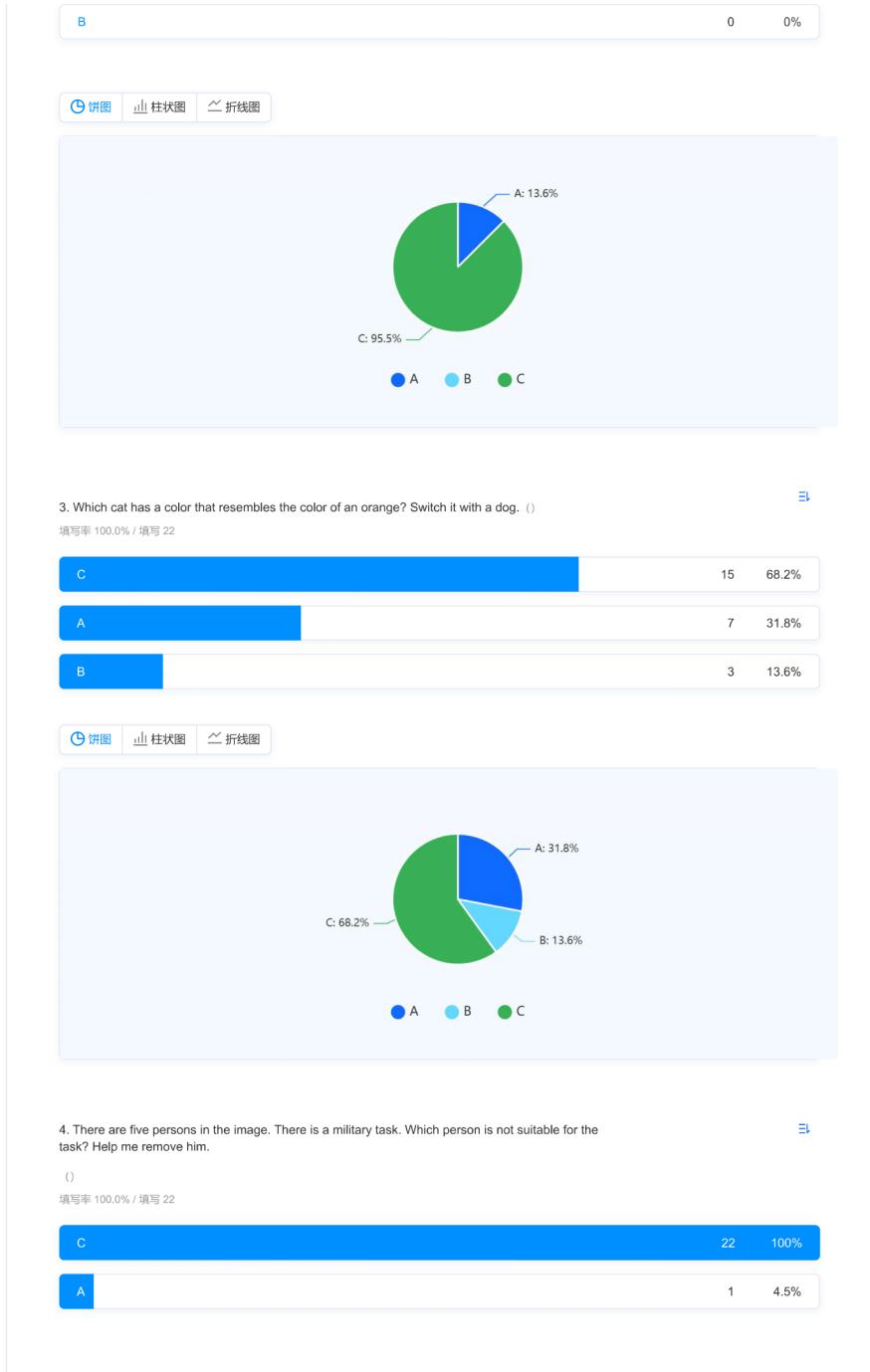


Figure A6: User Study Snapshot: Page 2

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

B 0 0%

饼图 柱状图 折线图

Category	Count	Percentage
A	2	9.1%
B	0	0%
C	22	100%

5. Skiing is a dangerous sport. In the picture, someone is not suitable for this sport. Please delete that person. ()

填写率 100.0% / 填写 22

Category	Count	Percentage
C	22	100%
A	2	9.1%
B	0	0%

饼图 柱状图 折线图

Category	Count	Percentage
A	2	9.1%
B	0	0%
C	22	100%

6. In this picture, who has poor eyesight? Exchange him with a firefighter. ()

填写率 100.0% / 填写 22

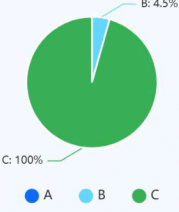
Category	Count	Percentage
C	22	100%
B	1	4.5%

Figure A7: User Study Snapshot: Page 3

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

A 0 0%

饼图 柱状图 折线图



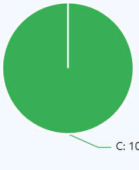
A	0	0%
B	4.5	4.5%
C	95.5	95.5%

7. Winter is coming, and animals need to rest. There is a sleepy bird in this picture. Help me exchange it with a frog. ()

填写率 100.0% / 填写 22

C	22	100%
A	0	0%
B	0	0%

饼图 柱状图 折线图



A	0	0%
B	0	0%
C	100	100%

8. Holiday is over. Someone needs to go to primary school. Let a dog replace that guy." ()

填写率 100.0% / 填写 22

C	22	100%
A	0	0%

Figure A8: User Study Snapshot: Page 4

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

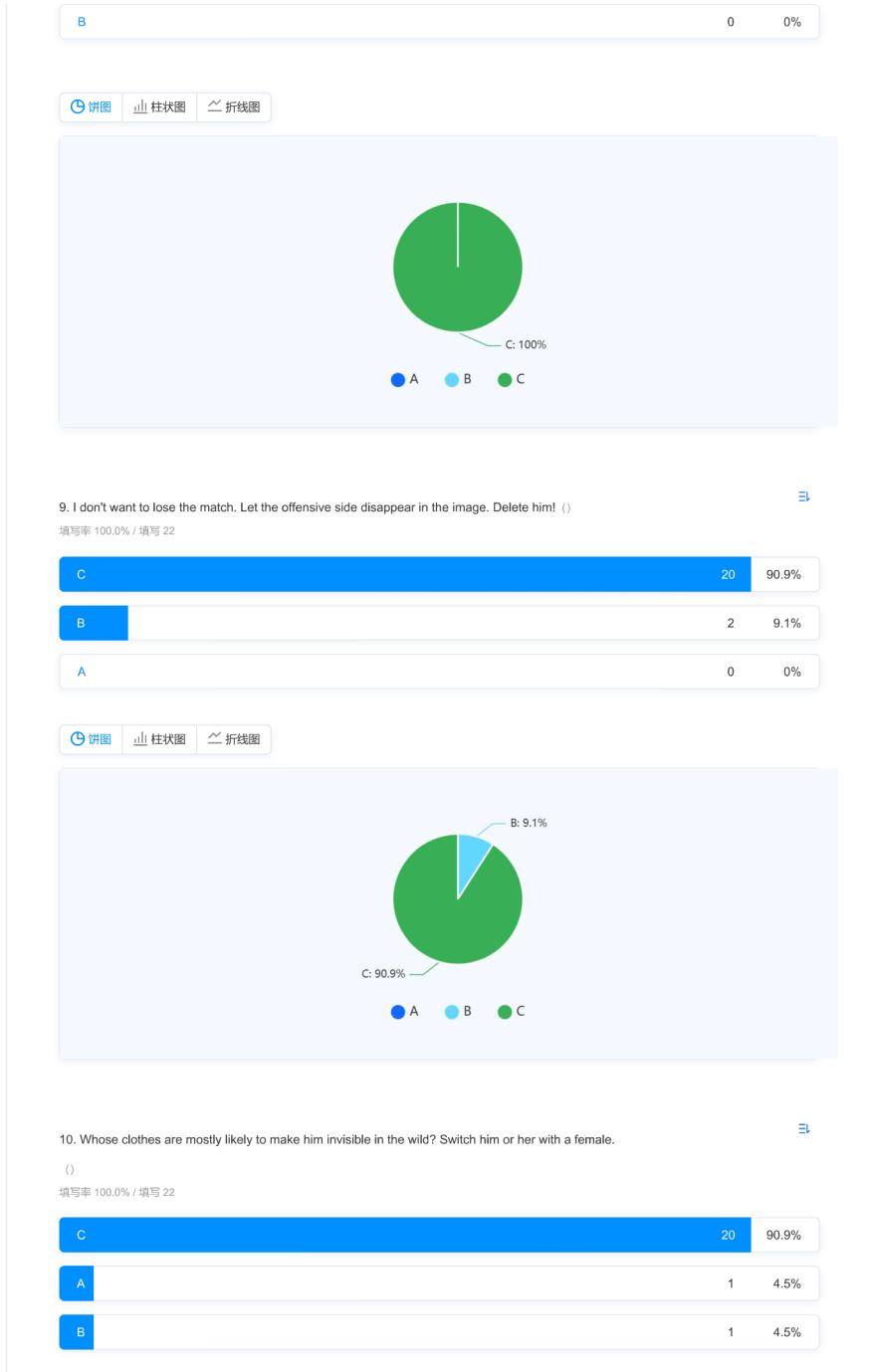


Figure A9: User Study Snapshot: Page 5

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
581
582
583
584
585
586
587
588
589
590
591
592
593

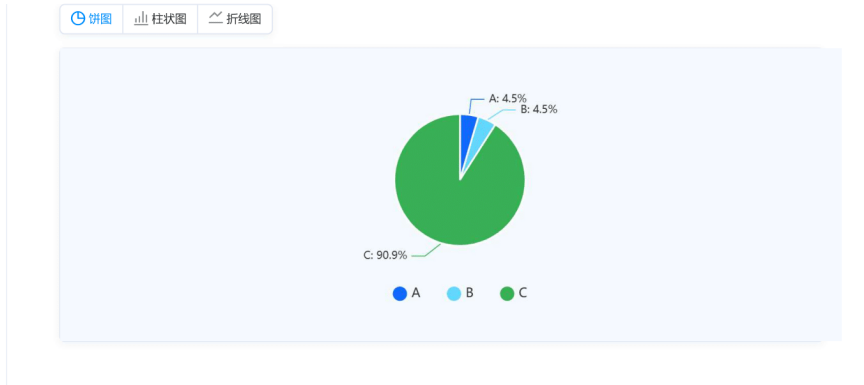


Figure A10: User Study Snapshot: Page 6

- Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 5
- Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 4
- Qingdong He, Xueqin Chen, Chaoyi Wang, Yanjie Pan, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xiangtai Li, and Jiangning Zhang. Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. *arXiv*, 2025. 1
- Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024. 1, 5
- Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: instruction reasoning dataset for advanced image editing. In *CVPRW*, 2024. 1
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023. 5
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4
- Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 4
- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL <https://arxiv.org/abs/2409.11340>. 4, 5
- Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. *arXiv preprint arXiv:2405.14785*, 2024. 1
- Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 1

594 Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi
595 Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking
596 reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025. 1
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