

Supplementary for “ADIEE: Automatic Dataset Creation and Scorer for Instruction-Guided Image Editing Evaluation”

Sherry X. Chen^{*1}, Yi Wei², Luowei Zhou^{†2}, and Suren Kumar²

¹University of California, Santa Barbara

²AI Center-Mountain View, Samsung Electronics

xchen774@ucsb.edu

{yi.wei1, luowei.zhou, suren.kumar}@samsung.com

In this supplementary, we first compare our model with a baseline VLM in the input and target image caption generation task for prompt-guided image editing models used in our dataset creation approach (Sec. A). Then, we provide more details regarding the experiment we have run, including dataset creation and model implementation in Sec. B, more information about the benchmarks we have tested (Sec. C), and lastly additional results from our scorer and editing models in Sec. D

A. Caption Generation for Prompt-Guided Editing Model

As mentioned in the method section, we need to generate input and target image captions in order to generate training samples from prompt-guided image editing models. Table 4 shows captions generated from a baseline VLM and from our fine-tuned VLM with respect to images in Fig. 9. As we can see, our model produces captions whose differences are more aligned with the edit instructions.

B. Dataset and Implementation Details

In our training dataset, for samples generate using text-guided image editing models, we use the model configuration as specified in ImagenHub [16]. We also use the same configuration for our fine-tuned MagicBrush model [27] to ensure a fair comparison with the baseline model.

When using LLaMA-Factory [29] to fine-tune a Qwen2-VL-7B-Instruct [23] for prompt generation, we follow the existing supervised fine-tuning (SFT) setting in the provided example in their code-base, which uses LoRA rank 8, batch size 8, and learning rate 1e-4.

When training the evaluation scorer, we create several different question templates to prompt the VLM:

^{*}Work done during an internship at Samsung AI Center Mountain View

[†]Corresponding author

Edit instruction: Make the woman fatter.	
Baseline	A woman is holding a pink umbrella and smiling. A woman in a blue dress holds a pink umbrella, standing in front of a building with a blue sky in the background.
Ours	a woman holding a pink umbrella a fat woman holding a pink umbrella
Edit instruction: add a plane in the sky	
Baseline	Two skateboarders are performing tricks on a graffiti-covered ramp in an outdoor skate park. A group of people are skateboarding on a graffiti-covered ramp in a park, with an airplane flying overhead and a tall building in the background.
Ours	a group of people skateboarding down a ramp a group of people skateboarding down a ramp with a small airplane flying in the sky
Edit instruction: Replace the red wine with white wine.	
Baseline	A bottle of Cupcake wine is placed next to a glass of red wine on a glass table, with a cushioned chair in the background. A bottle of Cupcake wine and a partially filled wine glass sit on a glass table outdoors.
Ours	a glass of red wine and a bottle of wine on a table a glass of white wine and a bottle of wine on a table

Table 4: Input prompt (above the dashline) and target prompt (below the dashline) generated by a baseline VLM [23] and our model corresponding to the edit instruction and input/groundtruth (GT) images in Fig. 9. The difference between each pair of input and target prompt is in bold. As we can see, our model produces prompt pairs whose difference is much more closely aligned with the edit instructions compared to the baseline.



Figure 9: We use 9 text-guided image editing methods - CycleDiffusion (CycleDiff) [24], DiffEdit [5], Prompt-to-Prompt (Pr2Pr) [9], pix2pix-zero (P2P-0) [19], SDEdit [18], Text2LIVE (T2L) [3], InstructPix2Pix (I-P2P) [4], MagicBrush [27], and AURORA [14] - to generate samples (faces blocked for privacy concerns) with various editing quality as part of our evaluation training data along with the ground-truth (GT) outputs. For methods that need input and target prompts to perform edits, we trained a VLM to get these prompts as shown in Table 4.

- “Can you rate how successful the edit instruction [INSTRUCTION] has been executed from the first image to the second image with a score from 0 to 10?”
- “Please rate how successful the edit instruction [INSTRUCTION] has been executed from the first image to the second image with a score from 0 to 10.”
- “How successful the edit instruction [INSTRUCTION] has been executed from the first image to the second image? Please respond with a score from 0 to 10.”
- “How successful the edit instruction [INSTRUCTION] has been executed from the first image to the second image? Please output a score from 0 to 10.”

where “[INSTRUCTION]” is a place-holder that will be replaced by the actual edit instruction with respect to the input and edited images.

Similar, we also create a few answering templates:

- “It is [SCORE].”
- “Sure, [SCORE]”
- “Sure, it is [SCORE]”
- “Sure, the score is [SCORE]”
- “[SCORE]”,

where “[SCORE]” is a special token that will be decoded to the final evaluation score. During training, we randomly

select a question and an answering template for each sample so the scorer is not over-fitted on fixed prompts.

The scorer training loss consists of two parts. The first part is the auto-regressive cross-entropy loss for VLM generated text $\hat{\mathbf{y}}_{txt}$ with respect to the selected answering template \mathbf{y}_{txt} , which is defined in LISA [17] as:

$$\mathcal{L}_{txt} = \text{CE}(\hat{\mathbf{y}}_{txt}, \mathbf{y}_{txt}). \quad (5)$$

The second part is the loss between the predicted score \hat{s} and the ground-truth score s :

$$\mathcal{L}_{score} = \text{L1}(\hat{s}, s). \quad (6)$$

The total training loss is

$$\mathcal{L} = \lambda_{txt}\mathcal{L}_{txt} + \lambda_{score}\mathcal{L}_{score}, \quad (7)$$

where we set $\lambda_{txt} = 1$ and $\lambda_{score} = 10$.

To fine-tune MagicBrush [27] using our scorer as a reward model in the reward condition setting, we take all samples in our evaluation training set that corresponds to input images and the edit instructions in MagicBrush training set to re-label them with reward scores predicted by our scorer. These samples become the training set to fine-tune MagicBrush on.

To fine-tune the model in the reward feedback learning setting, we train the model with a weighted sum of the reward feedback learning loss \mathcal{L}_{reward} and the original diffusion model MSE loss \mathcal{L}_{pre} as discussed in ImageReward [25], where the total training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{pre} + \lambda_{reward}\mathcal{L}_{reward} \quad (8)$$

for $\lambda_{reward} = 0.001$ so the weighted reward learning loss is balanced with the MSE loss for stable training.

C. Benchmarks Details

ImagenHub consists of 179 text-guided image editing samples, each containing an input image, edit instruction, input and target prompts, and a ground-truth output. Each sample includes editing outputs from eight methods [24, 5, 4, 27, 9, 19, 18, 3], evaluated by three human raters following a quantized scoring scheme: 0, 0.5, and 1. Following VIEScore [15], we compute the Spearman correlation between human ratings and predicted scores per method, applying Fisher Z-transformation to obtain the average correlation. The inter-rater (Human-to-Human) Spearman correlation serves as the upper bound for evaluator performance.

GenAI-Bench comprises 919 samples, each with an input image, edit instruction, two edit outputs from different methods, and a human preference label indicating which output is preferred or if both are good/bad. As it uses images in ImagenHub, each output is paired with a ground-truth edit.

AURORA-Bench contains two parts: (1) 2,000 point-wise evaluation samples with input images, edit instructions, edit outputs, and human-averaged quality scores ranging from 0–2, which we refer to as AURORA-Bench (point-wise); (2) 1,600 pair-wise comparison samples in the same format as GenAI-Bench but only distinguishes between preferred and tied outputs, and we refer them as AURORA-Bench (pair-wise). Samples from AURORA-Bench are collected from multiple sources: MagicBrush [27], Action-Genome [10], Something-Something [7], Epic-Kitchen [6], Kubric [8], CLEVR [12], WhatsUp [13], and Emu-Edit [20].

D. Additional Results

We include additional qualitative comparisons in Table 6 and 7. Notably, we include ImagenHub and AURORA-Bench (point-wise) results from baseline methods under the 1-shot setting defined in VIEScore [15], where a single image editing evaluation example with ground-truth outputs is included in the prompt input. Additionally, we include more qualitative results to compare our scorer with propriety models for image editing evaluation (Fig. 10) as well as image editing comparisons between MagicBrush [27] and our fine-tuned editing model in Fig. 11.

Lastly, we take all the editing models listed on the GenAI-Arena leaderboard [11] and rank them based on the average score of outputs generated by each model with respect to ImagenHub samples, as assessed by our scorer. The resulting scores and ranking is shown in Tab. 5, where our ranking aligns closely with the one from GenAI-Arena, despite our scorer not being trained on samples from four of

Method	Avg score	Rank	GenAI-Arena rank
MagicBrush	6.15	1	1
CosXL Edit	5.74	2	4
UltraEdit	4.63	3	2
InstructPix2Pix	4.18	4	5
Plug-and-Play	3.70	5	6
InfEdit	3.40	6	3
CycleDiffusion	3.20	7	8
Prompt-to-Prompt	3.00	8	7
SDEdit	1.41	9	9
pix2pix-zero	0.71	10	10

Table 5: Editing models in GenAI-Arena leaderboard [11] ranked based on the average score of their outputs with respect to ImagenHub [16] assessed by our scorer. The resulting ranking is closely aligned with the GenAI-Arena ranking, even though our scorer has not been trained on samples from four of the ten models: CosXL Edit [2], UltraEdit [28], Plug-and-Play [22], and InfEdit [26].

	ImagenHub AURORA-Bench (point-wise)	
Human-to-Human	0.4184	-
CLIP-D	0.2117	0.3080
CLIP-T	0.1894	0.1847
CLIP-I	0.1261	-
DINO-I	0.0441	-
GPT-4o	0.3821	0.4038
GPT-4o (1 shot)	0.3438	0.4779
Gemini-Pro 1.5	0.2728	0.1052
Gemini-Pro 1.5 (1 shot)	0.2648	0.2315
LLaVA	0.0273	0.0073
LLaVA (1 shot)	0.0258	-0.0110
LLaVA-NeXT	0.0356	-0.0491
LLaVA-NeXT (1 shot)	0.0468	0.0130
LLaVA-OneVision	0.0829	0.0555
LLaVA-OneVision (1 shot)	0.3225	0.0896
Qwen-VL	0.0404	0.0118
Qwen-VL (1 shot)	0.0037	0.0357
Qwen2-VL	0.1445	0.1783
Qwen2-VL (1 shot)	0.0914	0.1421
Qwen2.5-VL	0.1859	0.2351
Qwen2.5-VL (1 shot)	0.3467	0.2867
Phi3.5-vision-instruct	0.1126	-0.0107
Phi3.5-vision-instruct (1 shot)	0.2605	0.0381
Pixtral	0.0123	-0.0005
Pixtral (1 shot)	0.0243	-0.0005
BLIP-2	0.0378	-0.0003
BLIP-2 (1 shot)	-0.0085	0.0011
InstructBLIP	0.0212	-0.0351
Fuyu	0.0206	-0.0044
CogVLM	-0.0288	0.0199
OpenFlamingo	-0.0577	0.0065
ADIEE (Ours)	0.3450	0.4734

Table 6: Correlations of predicted scores with human ratings.

the ten models: CosXL Edit [2], UltraEdit [28], Plug-and-Play [22], and InfEdit [26].



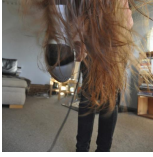

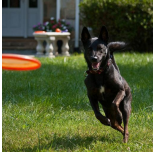
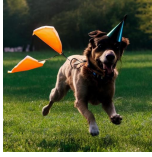


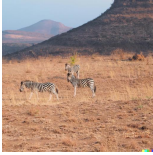
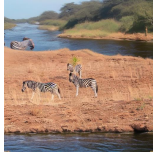
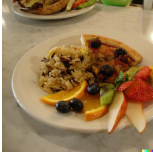
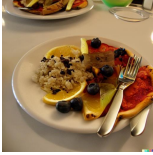
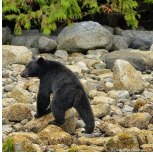
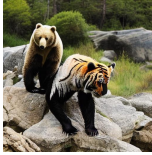

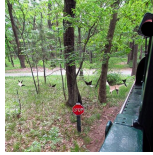




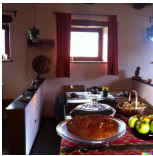


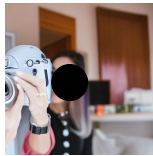
























Input	Output	Input	Output	Input	Output	Input	Output
							
“He should be eating a watermelon”		“turn her hair white”		“put a party hat on the dog”		“What if the man had a hat?”	
GT: 0.0 Ours: 0.94		GT: 0.0 Ours: 2.06		GT: 0.0 Ours: 2.71		GT: 2.36 Ours: 5.39	
GPT-4o: 1.41 Gemini: 3.87		GPT-4o: 4.90 Gemini: 5.20		GPT-4o: 5.65 Gemini: 8.37		GPT-4o: 6.0 Gemini: 7.75	
							
“put the zebras next to a river”		“There should be some cutlery on the table.”		“put a robot tiger next to the bear”		“Add a deer on the grass.”	
GT: 5.0 Ours: 6.09		GT: 6.67 Ours: 5.94		GT: 0.0 Ours: 0.23		GT: 3.33 Ours: 3.67	
GPT-4o: 3.74 Gemini: 7.07		GPT-4o: 4.90 Gemini: 4.90		GPT-4o: 2.45 Gemini: 2.82		GPT-4o: 0.0 Gemini: 0.0	
							
“make it a pepperoni pizza”		“A dog should be near the sheep.”		“Make the cake a chocolate cake”		“make the woman hold a camera”	
GT: 0.0 Ours: 1.66		GT: 3.33 Ours: 5.63		GT: 0.0 Ours: 2.37		GT: 0.0 Ours: 1.17	
GPT-4o: 4.90 Gemini: 6.71		GPT-4o: 5.66 Gemini: 6.32		GPT-4o: 3.46 Gemini: 3.46		GPT-4o: 2.65 Gemini: 7.07	
							
“turn the remote into a pizza”		“replace the donuts with fruits”		“Let the blueberry cake be topped with chocolate syrup.”		“Let’s add a cat on the roof.”	
GT: 0.0 Ours: 1.29		GT: 9.02 Ours: 6.76		GT: 7.07 Ours: 6.41		GT: 3.33 Ours: 4.41	
GPT-4o: 2.45 Gemini: 5.20		GPT-4o: 6.0 Gemini: 6.0		GPT-4o: 2.83 Gemini: 0.0		GPT-4o: 0.0 Gemini: 6.93	
							
“edit the background by removing the museum and placing a castle”		“remove the table and add an aquarium”		“let the kid sleep”		“Have a gorilla sit at the dinner table.”	
GT: 6.38 Ours: 6.09		GT: 5.0 Ours: 4.49		GT: 0.0 Ours: 1.45		GT: 1.67 Ours: 5.47	
GPT-4o: 5.66 Gemini: 5.66		GPT-4o: 0.0 Gemini: 0.0		GPT-4o: 3.16 Gemini: 7.75		GPT-4o: 5.66 Gemini: 6.32	
							
“Put a wooden floor on the kitchen.”		“put strawberry on the plate”		“Put a rat on the counter.”		“let it be a bullet train”	
GT: 10.0 Ours: 8.59		GT: 6.38 Ours: 4.20		GT: 5.69 Ours: 5.63		GT: 0.0 Ours: 2.97	
GPT-4o: 8.49 Gemini: 7.75		GPT-4o: 3.0 Gemini: 0.0		GPT-4o: 6.71 Gemini: 7.07		GPT-4o: 5.92 Gemini: 6.48	

Figure 10: More evaluation examples (faces are blocked due to privacy concerns) from GPT-4o [1], Gemini-Pro 1.5 (Gemini) [21], and our method on ImagenHub [16], where the ground-truth (GT) scores are presented below edit instructions.

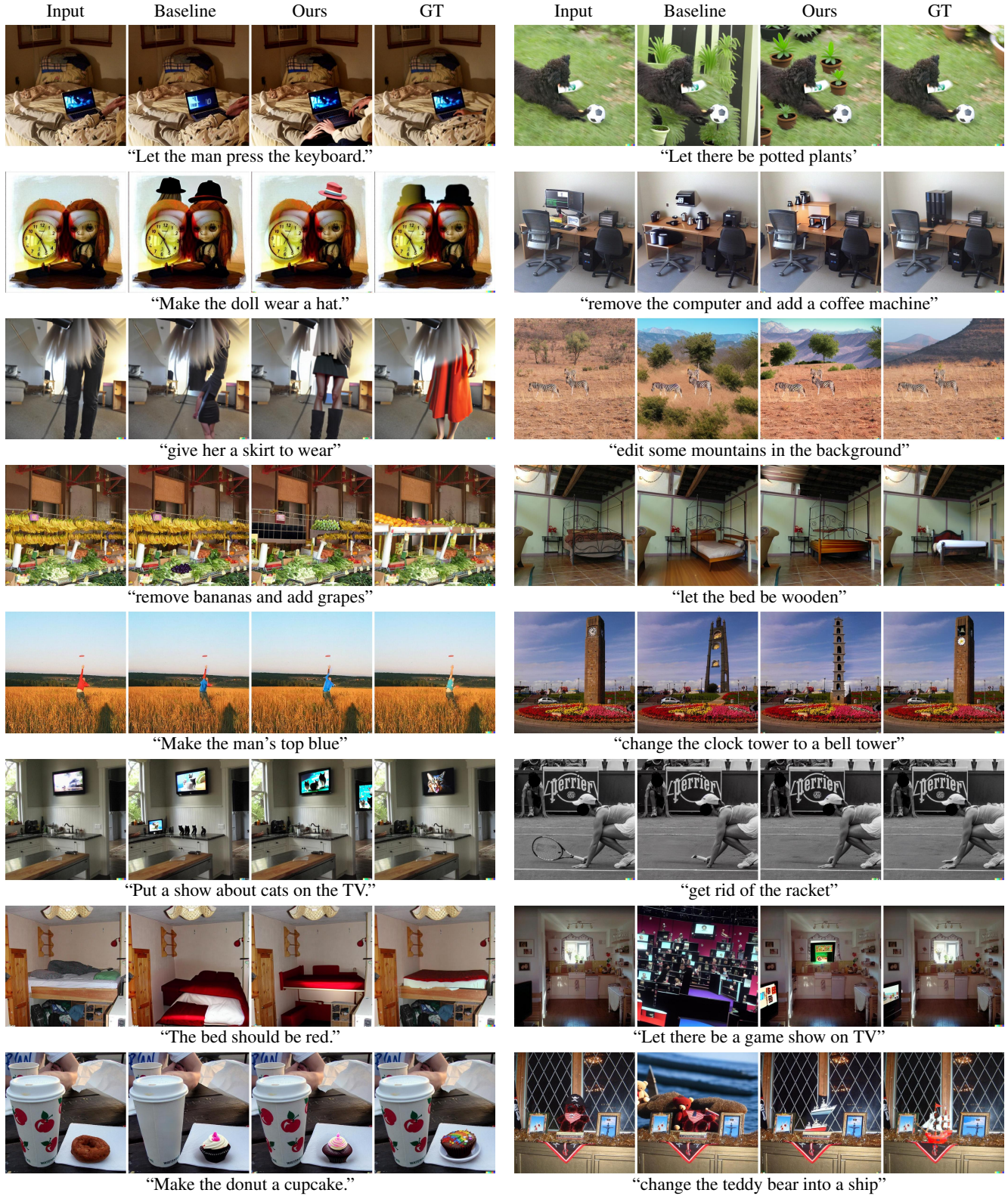


Figure 11: More mage editing examples (faces are blocked due to privacy concerns) from the MagicBrush editing model [27] and our fine-tuned model using our image editing scorer as a reward model.

	GenAI-Bench	AURORA-Bench (pair-wise)
random	25.90	33.43
CLIP-D	43.09	31.63
CLIP-T	39.39	42.93
CLIP-I	38.96	-
DINO-I	36.78	-
GPT-4o	53.54	50.81
Gemini-Pro 1.5	55.93	28.13
LLaVA	26.12	27.50
LLaVA-NeXT	25.35	27.19
LLaVA-OneVision	22.31	33.25
LLaVA-OneVision (1 shot)	22.31	33.25
Qwen-VL	14.91	12.69
Qwen2-VL	26.12	27.38
Qwen2-VL (1 shot)	26.12	27.38
Qwen2.5-VL	32.10	30.69
Qwen2.5-VL (1 shot)	0.0	31.31
Phi3.5-vision-instruct	21.87	32.25
Phi3.5-vision-instruct (1 shot)	21.87	32.25
Pixtral	26.12	27.38
Pixtral (1 shot)	26.12	27.38
BLIP-2	26.01	26.25
InstructBLIP	19.80	16.69
Fuyu	0.0	0.0
CogVLM	0.0	0.0
OpenFlamingo	0.0	0.0
ADIEE (Ours)	59.41	52.88

Table 7: Accuracy of predicted comparison labels with human preference.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4
- [2] Stability AI. Cos stable diffusion xl 1.0 and cos stable diffusion xl 1.0 edit, 2024. <https://huggingface.co/stabilityai/cosxl>. 3
- [3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 2, 3
- [4] 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*, pages 18392–18402, 2023. 2, 3
- [5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2, 3
- [6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021. 3
- [7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3
- [8] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanaprasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 3
- [9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3
- [10] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10236–10247, 2020. 3
- [11] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhui Chen. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*, 2024. 3
- [12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3
- [13] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023. 3
- [14] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning Action and Reasoning-Centric Image Editing from Videos and Simulations. In *NeurIPS*, 2024. Spotlight Paper. 2
- [15] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. 3
- [16] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenhui Zhuang, and Wenhui Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3, 4
- [17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

- [18] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3
- [19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3
- [20] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 3
- [21] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4
- [22] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play diffusion features for text-driven image-to-image translation. pages 1921–1930, 2023. 3
- [23] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [24] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023. 2, 3
- [25] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [26] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. 2024. 3
- [27] 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, 2024. 1, 2, 3, 5
- [28] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Ruijie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024. 3
- [29] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1