WeGen: A Unified Model for Interactive Multimodal Generation as We Chat (Supplementary Material)

Zhipeng Huang[1‡*] Shaobin Zhuang[2‡*] Canmiao Fu[3] Binxin Yang[3] Ying Zhang[3] Chong Sun[3]
Zhizheng Zhang[5†] Yali Wang[4†] Chen Li[3] Zheng-Jun Zha[1]
[1]University of Science and Technology of China [2]Shanghai Jiao Tong University
[3]WeChat Vision, Tencent Inc. [4]Chinese Academy of Sciences [5]Galbot

This supplementary material provides additional technical details (§1), extended experimental results (§2), and discusses limitations (§3) of WeGen.

1. More Details about WeGen

Visual Encoder-Decoder. As shown in Fig. 1, unlike VAE-based approaches, we adopt the CLIP model as our image encoder to leverage its semantic extraction capabilities, enabling efficient text-visual joint modeling with significantly reduced training cost and data requirements (Table 1 in the main paper). However, CLIP encoders often struggle with preserving fine-grained visual details. As discussed in the main paper, we observe that larger CLIP models better maintain visual details while preserving semantic extraction. Based on this, we employ a pretrained EVA-CLIP [18](4.9B) as our image encoder. Through bicubic interpolation of position embeddings, we extend the encoder to process 448×448 inputs instead of its original 224×224 resolution. The encoder outputs 16×16×1792 feature maps, which are pooled into a 64×1792 sequence, preserving both semantic information and visual details. For the decoder, we fully fine-tune SDXL's UNet weights, using a learning rate of 5e-4 with cosine scheduling and classifier-free guidance training by randomly drop 10% input image features. As shown in Figure 2, this configuration achieves superior reconstruction quality compared to existing methods.

Multi-modal Feature Modeling. As shown in Fig. 1, we adopt an autoregressive approach for visual feature modeling. Unlike parallel generation methods [2–4] that simultaneously predict all visual features from fixed placeholder tokens (*e.g.*<img1> to <img64>), our approach generates features sequentially with explicit dependencies:

$$P(x|c) = \prod_{i=1}^{64} P(x_i|x_{<i}, c) \tag{1}$$

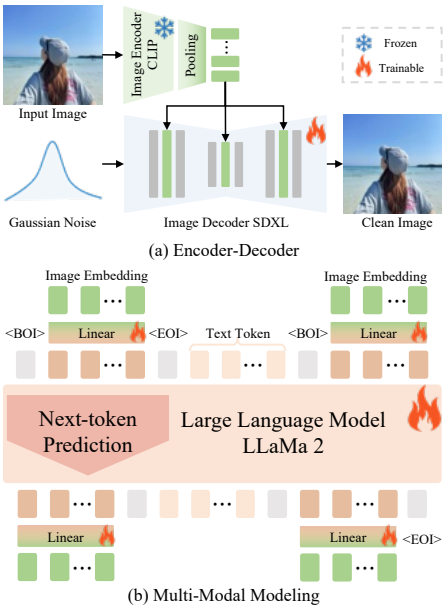

Figure 1. Detailed architecture of WeGen.

This explicit modeling of inter-feature dependencies enables our model to better capture holistic visual structure. Each term $P(x_i|x_{i-1}, ..., x_1, c)$ leverages previously generated features as context, rather than generating features in isolation ($P(x_i)$, $P(x_{i-1})$ ...). As shown in Figure 3, the quality difference becomes more evident with a fully fine-tuned UNet decoder. This is because when UNet focuses purely on decoding, generation quality heavily depends on MLLM's visual feature modeling, the parallel approach (left) shows blocking artifacts due to independent feature generation, while our autoregressive approach (right) maintains coherence through contextual generation. While parallel visual modeling approaches [2–4] rely on SDXL's pretrained weights and inherent generation capability to compensate for weaker MLLM visual feature modeling, this dependency on the original SDXL decoder limits the MLLM's

*Equal contribution ‡Work done as interns at WeChat
†Corresponding authors (zhangzz@galbot.com, yl.wang@siat.ac.cn)

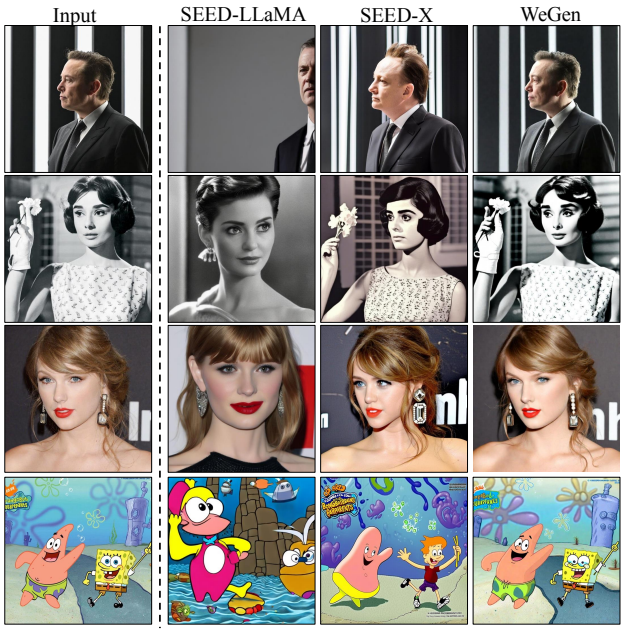

Figure 2. Qualitative comparison of reconstruction quality.

Task	Dataset
Reconstruction	Laion-COCO [8], Object365 [16], OpenImages [7]
Text2Image	**Laion-COCO-Recaption**(Ours), CapsFusion [22], JourneyDB [17]
Subject-Driven	GrIT [14], **DIIC**(Ours)
Restoration	Laion-COCO [8](Self-Aug), MultiGen-20M [15]
Editing	SEED-Edit [4], MagicBrush [23]
Condition Gen	MultiGen-20M [15], HR-VITON [9]
Style Transfer	StyleBooth [5], MultiGen-20M [15]
Understanding	**Laion-COCO-Recaption**(Ours), LLaVA-150K [10], LLaVAR [24], ScienceQA [12]

Table 1. Overview of training datasets.

fine-grained control over generation and editing tasks, making it challenging to achieve a truly unified visual design copilot.

Dataset Details Table 1 presents a comprehensive overview of the diverse datasets used for training our model. Our training leverages two primary datasets: (1) DIIC, containing 35M high-resolution frames with an average of 4.9 instances per frame and detailed captions (mean length 25.4 tokens); (2) Laion-COCO-Recaption, comprising 600M image-text pairs, each paired with both a concise caption (mean 10.2 tokens) and its expanded description (mean 79.6 tokens).

2. Additional Evaluation Results

Multi-Subject Generation Benchmark. We construct a multi-subject generation benchmark using CelebA-HQ [26]

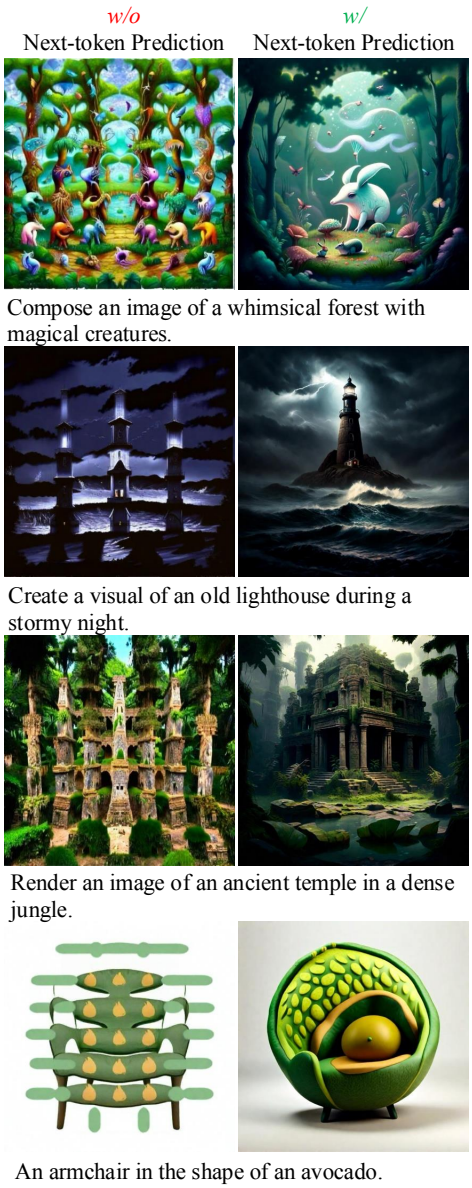

Figure 3. Visualization of feature modeling results. Left: parallel generation showing blocking artifacts. Right: our autoregressive generation producing more coherent visual features.

dataset, containing 2000 test cases with GPT-4 generated interaction prompts. Each case includes 2-3 reference faces. We evaluate using CLIP-T for text-image alignment, CLIP-I, DINO and face similarity[1] between reference and generated faces for identity preservation. As shown in Table 3, WeGen achieves superior performance across all metrics.

Understanding Capabilities. As shown in Table 4, while our primary focus is on unified visual generation for a de-

[1]Using face_recognition library (https://github.com/ageitgey/face_recognition)

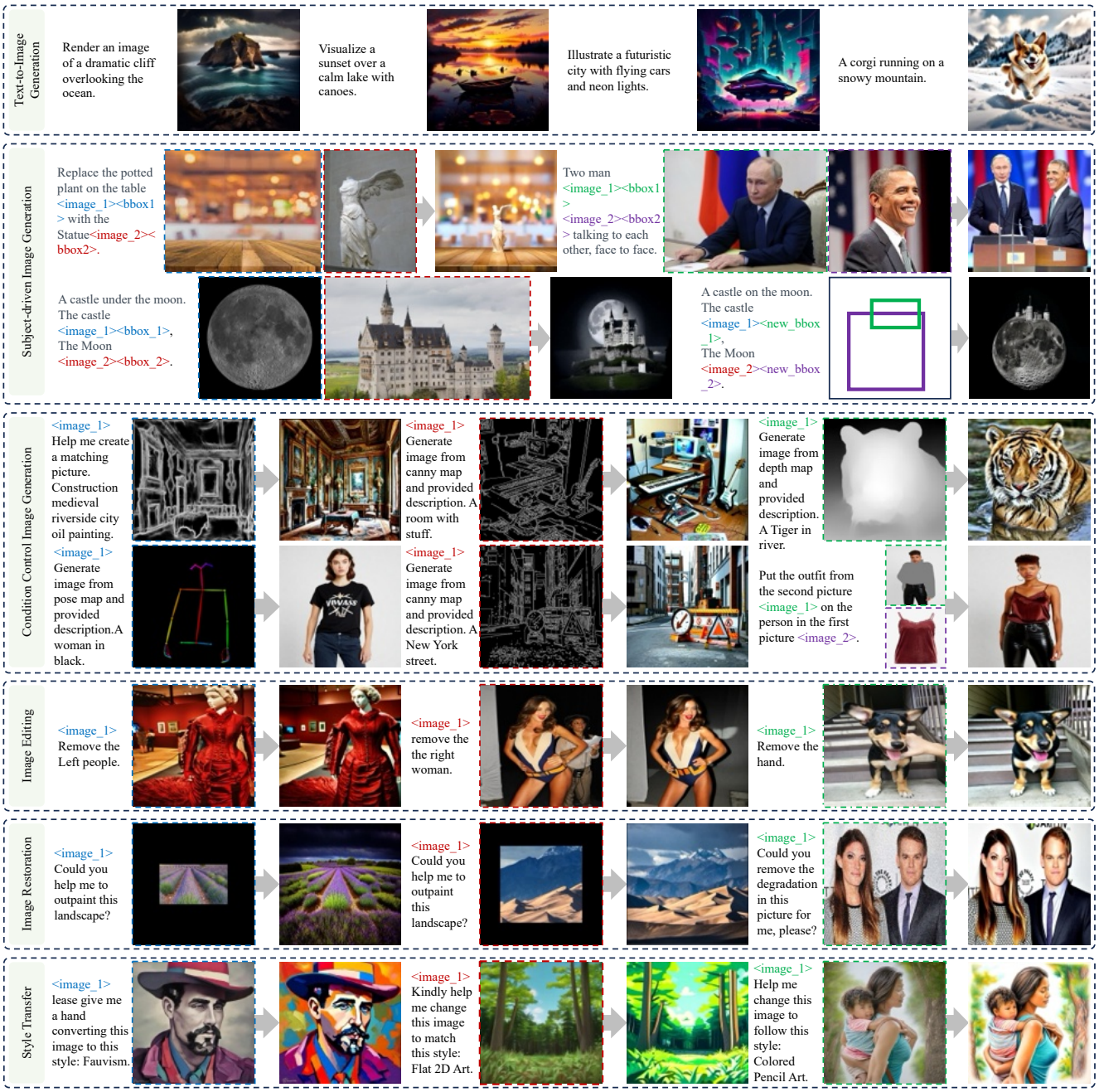

Figure 4. Extended case studies demonstrating WeGen's diverse capabilities across multiple visual generation tasks.

sign copilot, WeGen still achieves superior understanding performance[2] among unified models and maintains comparable results with understanding-only models across various visual understanding benchmarks.

Extended Case Studies. Figure 4 presents additional examples showcasing WeGen's capabilities across diverse tasks.

3. Limitations and Discussions

As shown in Figure 5, our approach exhibits degraded instance-level consistency when handling multiple reference images. While performing well with 2-3 references, the identity preservation deteriorates as reference number increases.

[2]All benchmarks are evaluated using VLMEvalKit (https://github.com/open-compass/VLMEvalKit)

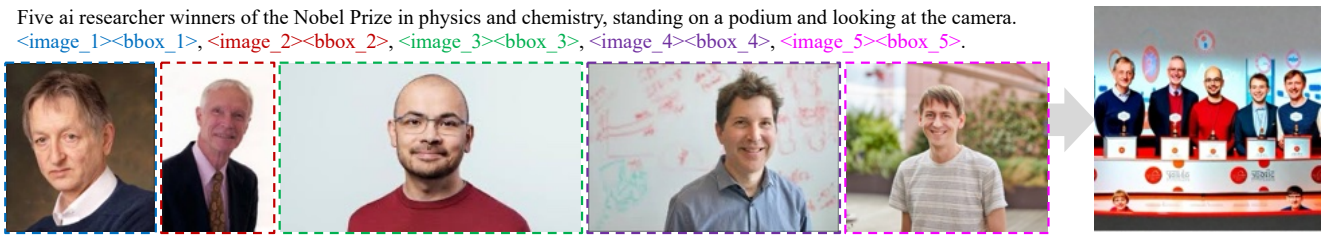

Figure 5. Failure cases with increasing number of reference images.

Configuration	Visual Decoding	Stage 1	Stage 2
Optimizer	AdamW		
Adam (β_1, β_2, ε)	$(0.9, 0.999, 10^{-8})$	$(0.9, 0.95, 10^{-6})$	
Peak LR	5×10^{-4}	5×10^{-4}	1×10^{-4}
LR schedule	cosine decay		
Gradient clip	1.0	5.0	
Training steps	5k	15k	5k
Warmup steps	1000		
batch size	4096	2048	
precision	`bfloat16`		

Table 2. Training hyperparameters across different stages.

Method	DINO (↑)	CLIP-I (↑)	Face Sim. (↑)	CLIP-T (↑)
Kosmos-G	0.583	0.712	19.1	0.285
Emu2	0.773	0.801	30.4	0.294
SEED-X	0.664	0.709	20.8	0.291
WeGen (Ours)	**0.803**	**0.845**	**52.4**	**0.294**

Table 3. Performance comparison on multi-subject generation benchmark. Face Sim. denotes face similarity.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 5

[2] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 1

[3] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[4] 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. 1, 2

[5] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 2

[6] Laurengon etc. Hugo. Idefics: Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023. 5

[7] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 2

[8] LAION. LAION-COCO: 600m synthetic captions from LAION-2B-en. https://laion.ai/blog/laion-coco/, 2023. Accessed: 2024-03-20. 2

[9] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2023. 2, 5

[11] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 5

[12] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[13] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 5

[14] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 5

[15] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 2

Type	Models	LLM Params	MMMU (↑)	Hallusion (↑)	MME (↑)	MMStar (↑)	MMT (↑)	OCRBench (↑)	ScienceQA (↑)	MMVet (↑)
Und.	MiniGPT4 [25]	7B	23.6	31.9	1047.4	16.3	16.5	172	39.6	15.6
	Kosmos-2 [14]	2B	23.7	19.8	721.1	24.9	25.5	244	32.7	23.7
	Idefics [6]	9B	18.4	27.3	1177.3	21.6	45.3	252	53.5	30.0
	LLaVA [10]	7B	34.1	21.6	28.3	27.1	1075.5	269	61.8	28.3
	Qwen-VL [1]	7B	29.6	29.9	482.7	32.5	42.9	127	61.1	13.0
	Emu2-Chat [19]	33B	40.7	29.5	1678.0	40.7	-	436	68.2	31.0
Und. & Gen.	Kosmos-G [13]	1.9B	14.8	20.4	104.3	18.4	18.3	109	29.6	11.3
	Chameleon-7b [20]	7B	22.4	17.1	202.7	31.1	23.9	5	46.8	8.3
	Gemini-Nano-1 [21]	1.8B	26.3	-	-	-	-	-	-	-
	LWM [11]	7B	-	-	-	-	-	-	-	9.6
	WeGen (Ours)	7B	**26.6**	**30.4**	**447.4**	**27.5**	**28.4**	**345**	**63.1**	**25.4**

Table 4. Performance comparison on visual understanding benchmarks. Und.: Understanding-only models; Und. & Gen.: Unified models with both understanding and generation capabilities.

[16] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[17] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[18] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[19] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 5

[20] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 5

[21] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[22] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. 2

[23] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[24] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[25] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5

[26] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 2