

MultiBanana: A Challenging Benchmark for Multi-Reference Text-to-Image Generation

Yuta Oshima^{1,*} Daiki Miyake^{1,*} Kohsei Matsutani¹ Yusuke Iwasawa¹
Masahiro Suzuki¹ Yutaka Matsuo¹ Hiroki Furuta^{2,†}

¹The University of Tokyo ²Google DeepMind

{yuta.oshima, daiki.miyake}@weblab.t.u-tokyo.ac.jp

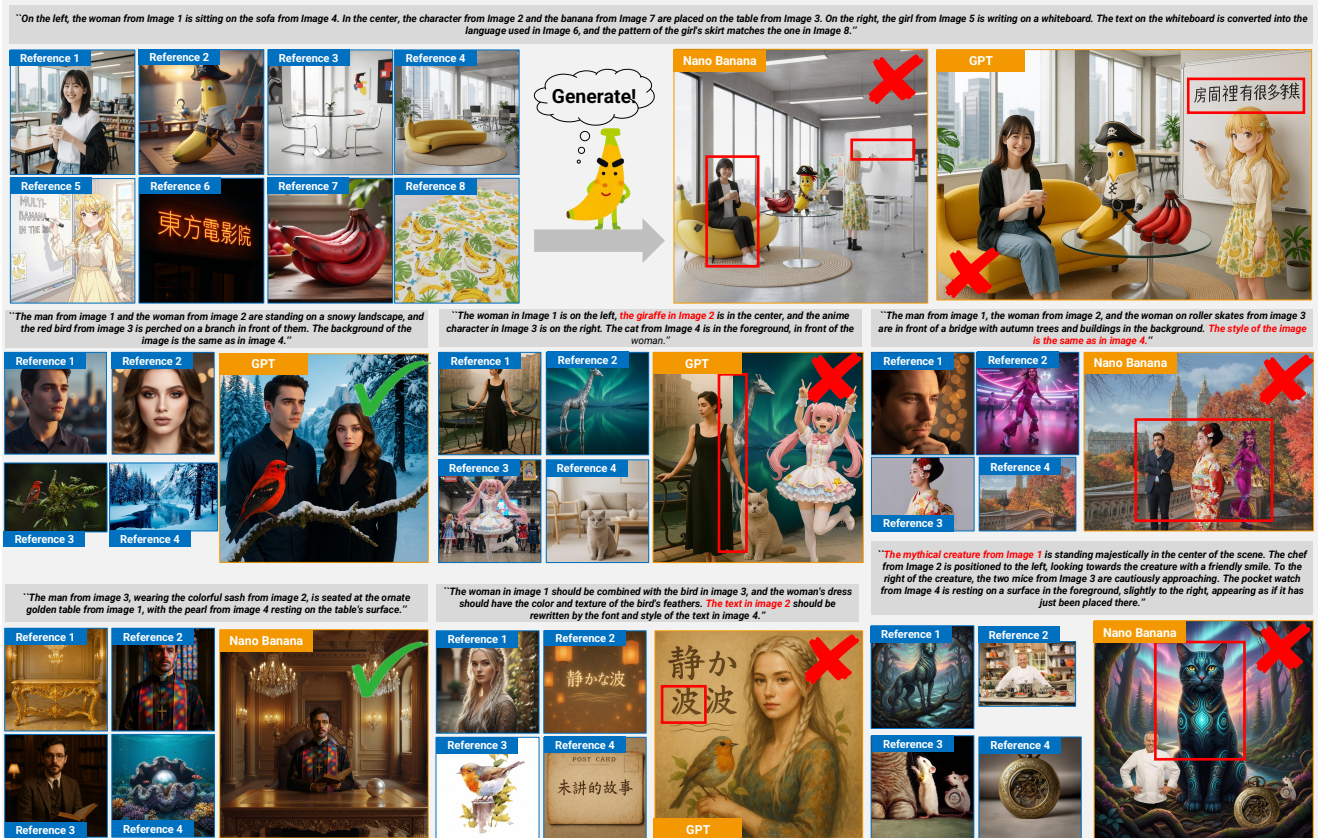


Figure 1. The overview of **MultiBanana**. MultiBanana broadly covers problems specific to multi-reference settings, including varying the number of references (top row), domain and scale mismatches among references (two on the left in the middle row), multilingual text rendering (center in the bottom row), and the presence of rare concepts (right in the bottom row).

Abstract

Recent text-to-image generation models have acquired the ability of multi-reference generation and editing; that is, to inherit the appearance of subjects from multiple reference images and re-render them in new contexts. However,

existing benchmark datasets often focus on generation using a single or a few reference images, which prevents us from measuring progress in model performance or identifying weaknesses when following instructions with a larger number of references. In addition, their task definitions are still vague, limited to axes such as “what to edit” or “how many references are given”, and therefore fail to capture the challenges inherent in combining heterogeneous refer-

*Equal contribution

†Work done as an advisory role only.

ences. To address this gap, we introduce **MultiBanana**, which is designed to assess the edge of model capabilities by widely covering problems specific to multi-reference settings: (1) varying the number of references (up to 8), (2) domain mismatch among references (e.g., photo vs. anime), (3) scale mismatch between reference and target scenes, (4) references containing rare concepts (e.g., a red banana), and (5) multilingual textual references for rendering. Our analysis among a variety of text-to-image models reveals their respective performances, typical failure modes, and areas for improvement. MultiBanana is released as an open benchmark to push the boundaries and establish a standardized basis for fair comparison in multi-reference image generation. Our data and code are available at <https://github.com/matsuolab/multibanana>.

1. Introduction

Recent advances in image generation, driven by large-scale multimodal LLM backbones, have enabled unprecedented levels of instruction-following [7, 13, 22, 56, 79–81, 83–85]. Among these developments, multi-reference image generation has emerged as a notable capability. When a user provides multiple reference images, the model can inherit the subject’s appearance and re-render it in a new context while faithfully preserving identity. This shift moves image generation beyond single-reference conditioning toward greater controllability, driving increasing demand in industrial domains such as content production [66, 78, 90], advertising [28, 33, 51], and fashion design [11, 16, 98], where reference-guided generation is essential.

Despite this progress, we lack a benchmark to evaluate how well models follow instructions that involve multiple references under varying conditions. As shown in Table 1, existing benchmarks [3, 30, 48, 66, 69, 77, 80, 83, 91–93] suffer from two fundamental limitations. First, existing benchmarks limit the number of references to a narrow range (typically 1–4) and do not examine how well models perform when following instructions with a larger number of references. Second, their task definitions are still vague, limited to axes such as “what to edit” or “how many references are given”, and therefore fail to capture the challenges inherent in combining heterogeneous references.

To address these limitations, we introduce a new evaluation framework that is grounded in the core challenges of multi-reference generation. Our benchmark incorporates multiple orthogonal axes specific to the multi-reference regime, including broader coverage of the number of references (up to 8), domain gaps between references (e.g., photo–anime mixtures), mismatches in object scale between reference and target scenes, rare concepts in the references (e.g., a red banana), and multilingual textual rendering references (English, Chinese, and Japanese).

Using this benchmark, we analyze model capabilities, identify characteristic failure modes, and highlight remaining gaps. For example, our analysis reveals that as instruction difficulty increases (through more references, cross-domain mixtures, or rare concepts), closed-source models (e.g., Nano Banana) tend to incorporate all required subjects but suffer from compositional distortion due to overfitting to reference details, while open-source models (e.g., Qwen-Image) generate visually cleaner outputs but often omit some reference subjects altogether. We release MultiBanana as an open-source benchmark to establish a standard foundation for fair comparison and to accelerate progress in multi-reference image generation.

2. Related Works

2.1. Controllable Text-to-Image Generation

Diffusion models have achieved state-of-the-art performance in high-fidelity image synthesis [26, 70]. Large pre-trained diffusion models with textual conditioning, such as Stable Diffusion [15, 60, 65] and FLUX [41, 42], can generate novel images while preserving the identity depicted in the reference images. Building on these developments, approaches based on fine-tuning (e.g., DreamBooth [66]) and adapter-based methods (e.g., IP-Adapter [90]) have been explored to improve the controllability [4, 18, 34, 44, 50, 72]. Recent research on multimodal LLM-based image generation has gained momentum, aiming to handle multiple image generation tasks within a single model. For example, GPT-Image-1 [56] and Nano Banana [22] can jointly process text and image inputs, enabling integrated image generation and editing within a unified framework. Similarly, open-source models such as Qwen-Image [79] and FLUX.1-Kontext-dev [42] demonstrate high-quality, flexible image generation and editing. Beyond these, numerous open-source unified generative models continue to emerge [80, 83–85]. These advancements demonstrate that diffusion models are becoming flexible and adaptable across diverse conditional generation settings.

2.2. Benchmarks for Reference-Based Generation

Among reference-driven image generation tasks, image editing is the most widely recognized: it involves generating an edited image from a single reference image and a textual instruction describing how the edit should be performed. Representative benchmarks such as MagicBrush [93] and EMU-Edit [69] have extended task-specific evaluations but still rely heavily on similarity-based metrics. SmartEdit [30] supports editing of complex scenes, but does not sufficiently cover more general settings. Similarly, I2E-Bench [48] employs GPT-4o to provide human-aligned evaluations across a variety of editing tasks, yet it uses different metrics for each task and therefore fails to capture the

Table 1. Comparison among major benchmarks for reference-based image generation. Existing benchmarks do not provide systematic evaluation across diverse multi-reference conditions, support only a limited number of references, and fail to adequately account for important factors such as differences and compatibility among reference images. Our benchmark expands the upper limit on the number of references and introduces *difficult reference combinations*, including domain mismatches, scale mismatches, rare concepts, and multilingual prompts, thereby explicitly addressing challenges unique to multi-reference image generation.

Benchmark	#Size	#Tasks	#References	Difficult Reference Combination	Metrics
EditBench [77]	240	1	1	✗	CLIP [62]
EditVal [3]	648	13	1	✗	CLIP, VLM, manual
EmuEdit [69]	3055	7	1	✗	L1, CLIP, DINO [8]
MagicBrush [93]	1053	9	1	✗	L1, L2, CLIP, DINO
AnyEdit [92]	1250	25	1	✗	L1, CLIP, DINO
I2EBench [48]	2240	16	1	✗	GPT [54]
ImgEdit-Bench [91]	811	14	1	✗	GPT (3 dim.), Fake Det. [87]
DreamBooth [66]	75	1	1	✗	CLIP, DINO
OmniContext [80]	400	8	3	✗	GPT (3 dim.)
DreamOmni2 [83]	319	29	4	✗	Gemini [19], Doubao [5]
MultiBanana (Ours)	3769	36	8	✓	GPT, Gemini (5 dim.)

shared characteristics of editing as a whole. ImgEdit [91] has enabled evaluation of multi-turn editing tasks, in which image generation and editing are alternated interactively. One of the earliest benchmarks for reference-based image generation beyond instruction-based editing is the DreamBooth [66] dataset. However, such benchmarks are still limited to generation tasks conditioned on a single reference image and thus cannot handle more complex or compositional settings. Recently, advances demonstrated by GPT-Image-1 [56] and Nano Banana [22] have drawn increasing attention to multi-reference image generation, where multiple reference images are jointly used for generation. Benchmarks such as OmniContext [80] and DreamOmni2 [83] have been proposed to evaluate this task. Nevertheless, these benchmarks encompass only a small number of images and task types, thereby limiting the scope of the evaluation. Moreover, although designed for multi-reference generation, they do not account for the heterogeneous nature of reference images. This challenge is analogous to instruction following in LLMs [23, 24, 35, 40, 97], where performance degrades as the number of constraints grows; multi-reference image generation similarly requires models to satisfy all given references without neglecting any.

3. MultiBanana

3.1. Task Types

In reference-driven image generation, tasks can be categorized based on the number and relational structure of reference images. In this work, we follow the overall design philosophy of prior benchmarks such as DreamOmni2 [83], but extend the number of references up to 8 to enable a more fine-grained analysis of scalability. We also go beyond task-

category-based evaluation by incorporating axes that capture the intrinsic properties of the reference set itself, such as domain gaps between references, scale mismatches, rare concepts, and multilingual text rendering.

Single Reference. This is the simplest setting, corresponding to the standard image editing task. Given a single reference image and a textual instruction, the model generates an edited image that maintains semantic consistency and preserves the subject’s visual identity. This configuration follows traditional instruction-driven editing benchmarks such as MagicBrush [93] and EMU-Edit [69], which represent the foundation of reference-driven generation.

Two References – 11 Task Types. The two-reference setting provides the model with two reference images (typically a main subject and an auxiliary one) and a textual instruction. We define 11 editing task types, following prior work, while balancing feasibility and challenge: subject addition, subject replacement, background change, color modification, material modification, pose modification, hairstyle modification, makeup modification, tone transformation (lighting, season, weather, etc.), style transfer (oil painting, sketch, anime, etc.), and text correction (in-image text editing or replacement). These tasks span both local attribute manipulations and global transformations, evaluating a model’s ability to maintain visual coherence, subject identity, and contextual adaptability.

Multi References – Compositional Generation and Inter-Reference Relationships. The multi-reference setting uses 3–8 reference images, requiring the model to perform more compositional generation. We define four base task structures, each evaluated with {3, 4, 5, 6, 7, 8} reference images, resulting in 24 tasks:

- **X Objects:** Compose X objects from the reference set.

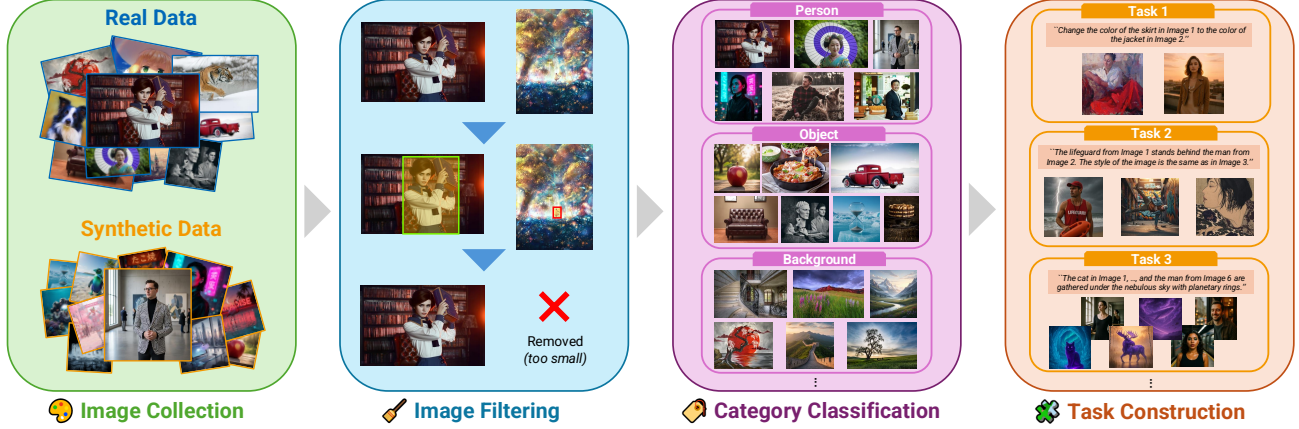


Figure 2. Construction pipeline for our benchmark, consisting of four stages: (1) collecting high-quality real and synthetic images, (2) filtering out inappropriate or low-quality samples, (3) performing hierarchical category classification, and (4) generating and validating editing instructions by Gemini and humans.

- **X-1 Objects + Background:** Place $X - 1$ objects within the background of another reference.
- **X-1 Objects + Local:** Apply a local change (specified by a reference) to one of the $X - 1$ objects.
- **X-1 Objects + Global:** Arrange $X - 1$ objects and modify the overall tone according to another reference.

This design enables us to assess whether a model can integrate multiple reference sources into a consistent scene.

Diversity Factors in Multi-Reference Generation. A central contribution of our benchmark is the explicit incorporation of inter-reference diversity factors, which have been largely overlooked in previous benchmarks such as DreamOmni2 [83]. We highlight the following dimensions:

- **Cross-domain diversity:** mixing references from different domains such as real photos, anime, CG, or sketches.
- **Scale and viewpoint differences:** variation in object size or camera perspective between references and outputs.
- **Rare concept references:** inclusion of uncommon concepts (e.g., “a red banana”).
- **Multilingual references:** assessing multilingual instruction understanding during rendering.

These factors go beyond prior benchmarks that focus on simple reference composition; our benchmark evaluates whether models can interpret relationships among heterogeneous references and integrate them into coherent outputs.

3.2. How to Construct MultiBanana

The construction pipeline is illustrated in Figure 2.

Image Collection. The reference images used in our benchmark consist of both real and synthetic data. All real images were extracted from the LAION-5B dataset [68]. To ensure high quality, we selected only images with an Aesthetic score [43] higher than 6.25 and a resolution greater than 512 pixels. However, the collected real images exhibited a distributional bias (Figure 3; left). Specifically, land-

scape scenes appeared frequently, whereas humans and objects were underrepresented. While landscape images can serve as useful background references, the scarcity of human and object references limits the diversity of editing tasks. To address this imbalance, we generated additional synthetic images under diverse conditions using Nano Banana [22] and GPT-Image-1 [56], and used them as supplementary reference images (Figure 3; right). This approach not only mitigated the categorical bias in the reference set but also expanded its coverage and diversity by incorporating both real and synthetic examples.

Image Filtering. Among the collected images, some contained small subjects that were difficult to use as references, or were inappropriate for reference use, such as images containing harmful content, tables, or corrupted data. To filter out such images and retain only those suitable for a reference-based image editing benchmark, we applied the following procedure. Following [91], we extracted images that contained large and clearly visible objects suitable for editing tasks. Object detection was performed using YOLOv12 [73] and SAM2 [64], while CLIP [62] was used to verify semantic consistency. Additionally, harmful or corrupted images were removed through a combination of automatic filtering with Gemini and manual inspection, resulting in the final curated set of real images.

Category Classification. To construct multi-reference image generation tasks, we categorized each reference image. For example, in a portrait transformation task that modifies a person’s hairstyle or makeup, both reference images must contain a person. Similarly, in replacement tasks involving humans or objects, both reference images must include either a person or an object, and for background replacement tasks, images categorized as background must be used. To satisfy such task-dependent conditions, we performed hierarchical image classification on a combined

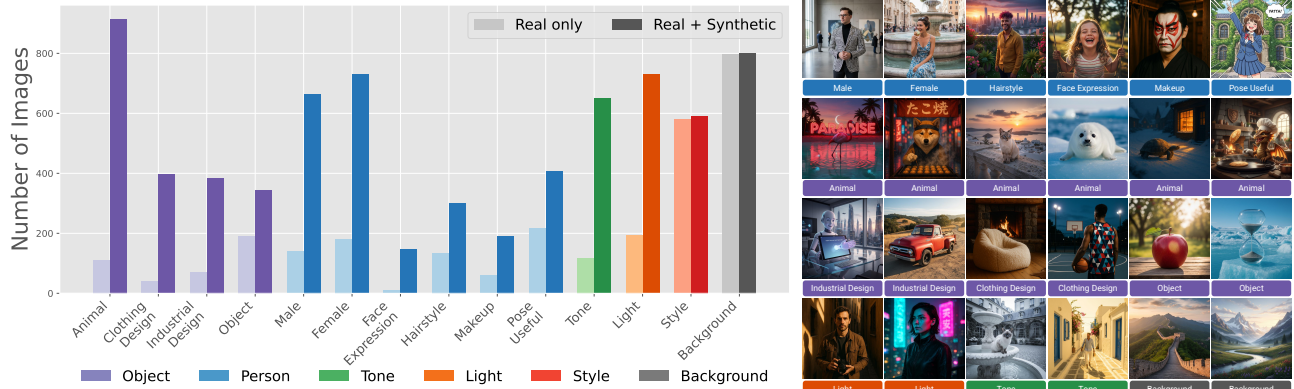


Figure 3. **(Left)** Comparison between the statistics of real data only and those after adding synthetic data. The original dataset was biased toward background images, with few person- and object-related samples. To correct this imbalance, we generated additional synthetic images using Nano Banana and GPT-Image-1, focusing on clear subjects such as people, animals, and objects. This results in a more balanced and comprehensive distribution of data. **(Right)** Examples of synthesized images in each category.

set of real and synthetic images. We first obtained automated multi-level labels using Gemini, then conducted human verification to ensure correct categorization. At the top level, we defined six major categories: *person*, *object*, *background*, *light*, *tone*, and *style*. Each of these top-level categories was further divided into more specific subcategories. For instance, under the *person* category, we included *male*, *female*, *hairstyle*, *makeup*, *pose useful*, and *face expression*; and under the *object* category, we defined *animal*, *clothing design*, *industrial design*, and *object*. This hierarchical labeling enabled systematic task construction.

Task Construction. The editing instructions were generated using Gemini. For each task, we manually selected suitable reference categories, randomly sampled images, and presented them to Gemini to generate corresponding editing instructions. Gemini then evaluated whether following each instruction would lead to visual breakdowns. For instance, in a replacement task, whether the replaced object would appear unnaturally suspended in mid-air. Among the generated candidates, we retained only the reference-instruction pairs that did not exhibit visual breakdowns and then asked Gemini and human annotators to further filter out tasks that were nonsensical or too trivial.

Difficult Reference Categorization. For the cross-domain category, we selected samples whose reference images spanned at least two domains (e.g., color photos, anime, paintings). The different-scale category included samples with both close-up and non-close-up views. The rare-concept category contained samples featuring animals with unusual colors or patterns, humans or animals with atypical makeup or clothing, or fictional creatures. The multilingual category included samples with text in two different languages, including English, Chinese, and Japanese.

3.3. Statistics in MultiBanana

Image Category Statistics. This section presents a statistical overview of both the real and synthetic source data used to construct our benchmark. The dataset is organized into six major categories and thirteen subcategories based on visual and semantic attributes.

In the real dataset, the largest major category is background, comprising 798 images, one-fourth of the total. These mainly consist of images without distinct subjects. The next-largest categories are person (741 images) and style (579 images), and together these three categories account for 70% of the dataset. In contrast, light (195 images), object (413 images), and tone (118 images) are small in number. Although these categories are useful as references for changing global conditions such as illumination and overall atmosphere, their sample sizes remain limited. Within the object category, there are 111 images of animals, 40 of clothing and textile design, 71 of industrial design, and 191 of miscellaneous objects, indicating that data focusing on specific materials or design patterns is scarce. Most images in the background, style, and tone categories lack clear subjects. Furthermore, images containing textual elements, such as logos or signs, are nearly absent. The person category is subdivided into finer attributes: face expression (10), female (181), male (139), hairstyle (135), makeup (59), and pose usefulness (217). While pose-related data account for the largest portion (around 30%), reference data capturing subtle characteristics such as facial expression or makeup remain limited, resulting in a skewed distribution.

To address this imbalance, synthetic data were generated, focusing primarily on images containing explicit subjects (e.g., persons, animals, and objects) and on those including text. The generation process was designed to intentionally expand the coverage of underrepresented subcategories and improve the overall balance among attributes.

Table 3. Average performance per model across different task categories (10-point scale; the higher the better). The average scores from Gemini 2.5 and GPT-5 are reported. Both open-source and closed-source models exhibit notably lower performance on background replacement tasks, regardless of the number of reference images.

Model	Single	Two	X Objects	X-1 + Local	X-1 + Global	X-1 + Background
Nano Banana	7.817	4.891	4.453	4.118	4.698	3.575
+ Agent (IPR)	7.606	5.030	4.433	4.030	4.496	3.767
GPT-Image-1	7.804	6.585	5.086	5.147	5.757	5.019
+ Agent (IPR)	7.820	6.808	5.258	5.419	5.775	5.284
Qwen-Image-Edit-2509	7.499	3.699	2.256	2.031	2.351	2.033
DreamOmni2	6.520	4.069	2.804	3.037	2.867	2.594
OmniGen2	5.919	3.442	3.256	3.598	3.369	3.022

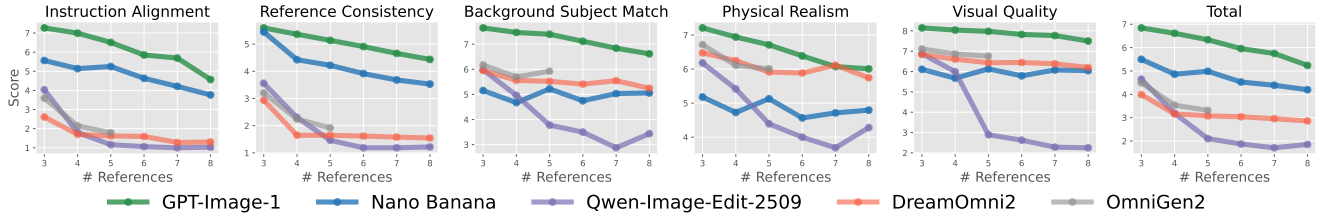


Figure 5. Changes in scores for each evaluation criterion when varying the number of reference images. Both open-source and closed-source models exhibit a general trend of decreasing scores across all metrics as the number of references increases.

reference images, we evaluate it only under the 5-reference setting. As described in Section 3.4, Gemini 2.5 [19] and GPT-5 [57] are used for evaluation. In the main paper, we report the average scores from Gemini 2.5 and GPT-5. See Appendix A for the results of Qwen3-VL as judge, and Appendix F for the detailed results and discussion.

4.1. Per-Task Evaluation

We compute the MultiBanana scores for each task type defined in Section 3.1, and present the results in Table 3. For the two-reference tasks, we report the average score across all 11 task variants. For the multi-reference tasks, we compute the average score for each task type (i.e., X Objects, X-1 Objects + Local, X-1 Objects + Global, X-1 Objects + Background) across $X = \{3, 4, 5, 6, 7, 8\}$.

Both open-source and closed-source models exhibit lower performance on background-replacement tasks, regardless of the number of reference images. The results also reveal clear performance gaps between closed-source and open-source models, even in relatively simple editing scenarios, such as single- and two-reference tasks.

4.2. Effect of the Number of References

We investigate how the MultiBanana score changes when varying the number of reference images $X = \{3, 4, 5, 6, 7, 8\}$ for each multi-reference generation task (Figure 5). Both open-source and closed-source models exhibit a general trend of decreasing Total Score with increasing number of references. However, the nature of this

degradation differs between the two model families.

Closed-source models demonstrate strong adherence to reference attributes and prompt instructions. As a result, even when the number of references grows, the decline in Instruction Alignment and Reference Consistency remains relatively mild. Nevertheless, because these models attempt to strictly satisfy all reference constraints, increasing the number of subjects often degrades overall visual quality, including overcrowded scenes and compositional inconsistencies (see Figure 1 and Figure 6; left).

In contrast, open-source models exhibit weaker adherence to prompts and references; even under the 3-reference setting, both Instruction Alignment and Reference Consistency remain low, and these scores deteriorate further as the number of references increases. Qualitative inspection also reveals that open models frequently ignore multiple subjects under high-reference conditions. However, their overall quality (background consistency, physical realism, visual quality) remains relatively stable, and they tend not to suffer from the severe compositional collapse observed in closed models (see Figure 6; right).

In summary, closed-source models succeed in incorporating all required subjects but often generate globally inconsistent or distorted scenes due to overfitting to fine-grained reference details. Open-source models, by contrast, tend to omit multiple reference subjects in complex scenarios but generate visually clean images with fewer structural failures. These observations highlight a clear trade-off between strict reference fidelity and holistic image coherence.

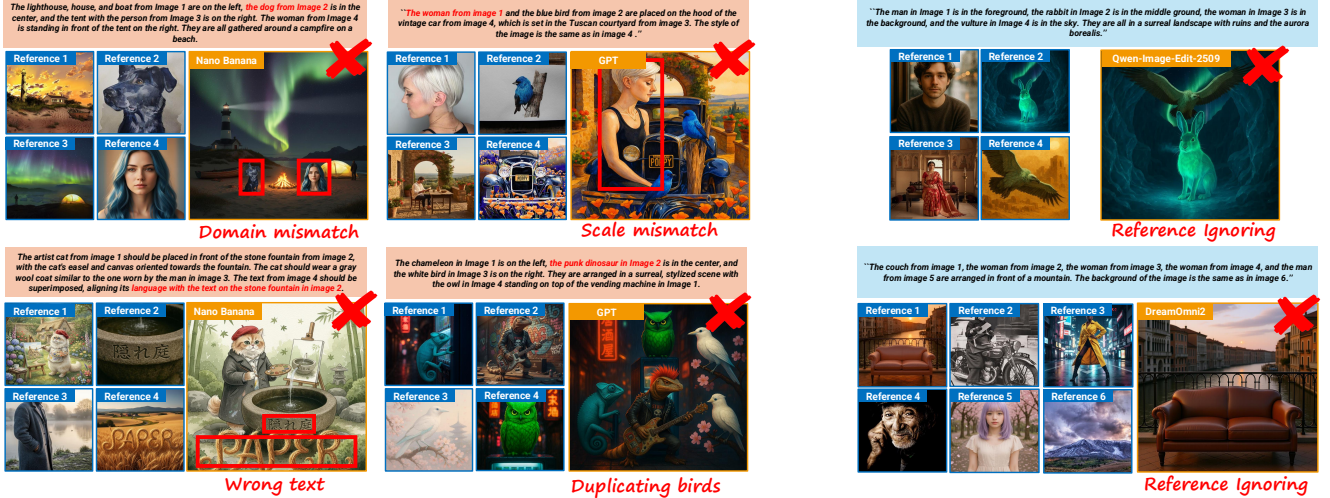


Figure 6. **(Left)** Failure cases of closed-source models under challenging conditions such as cross-domain references, scale mismatches, rare concepts, and multilingual text. **(Right)** Failure cases of open-source models under many-reference conditions. They tend to ignore some reference subjects entirely.

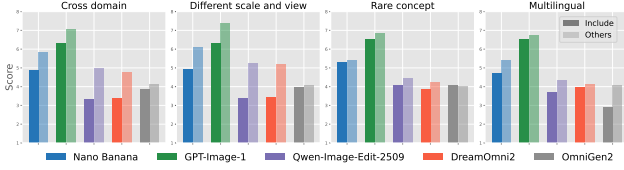


Figure 7. Results for difficult reference combinations. For cross-domain and different scale and view tasks, every model shows lower scores than tasks without such conditions.

4.3. Analysis of Difficult Reference Combinations

To analyze the challenging reference combinations described in Section 3.1, we compare MultiBanana scores between tasks containing these difficult combinations and those that do not, as shown in Figure 7. For cross-domain, different scale and view, rare-concept, and multilingual tasks, most models achieve lower scores than on tasks without such conditions. These results indicate that the difficult reference concepts focused on by our benchmark remain challenging, even for state-of-the-art closed models. This demonstrates that our benchmark effectively uncovers the current limitations of multi-reference image generation.

4.4. Reliability of VLM Judge

To assess the reliability of VLM-based evaluation, we measured the correlation between VLM and human ratings using three human evaluators on 32 samples from our benchmark (Table 4). Both GPT-5 and Gemini 2.5 show strong correlation with human judgments. This validates the use of VLM judges as a time- and cost-efficient proxy for human evaluation, as described in Section 3.4. We also confirmed that Qwen3-VL-8B-Instruct [2] is a reliable open-source alternative, thereby broadening accessibility for researchers without API access. See Appendix E for further discussion.

Table 4. Correlation between human and VLM judges.

Judge	Pearson r	Spearman r	Cohen’s κ
GPT-5	0.6936	0.7134	0.6135
Gemini 2.5	0.5726	0.6113	0.3908
Qwen3-VL	0.5457	0.5642	0.5235

4.5. Agentic Framework

To address the performance limitations on our challenging benchmark, we further construct three agentic framework baselines: Iterative Prompt Refinement (IPR), Context-Aware Feedback Generation (CAFG), and Selective Reference Adaptation (SRA). We then conduct experiments on our MultiBanana benchmark. We show the results of IPR in Table 3, where the generator creates images, and the planner rewrites the prompt over multiple steps based on these images. While GPT’s performance improved with the agent framework, Nano Banana showed no improvement and, in some cases, performance degraded due to information loss from the initial instruction. For details on the problem setting, ablations, and results, see Appendix F.7.

5. Conclusion

We constructed a comprehensive benchmark for multi-reference image generation, focusing on challenges specific to this setting, such as difficult reference combinations. Using it, we evaluated state-of-the-art open and closed models and found two notable failure modes as the number of references increased: models either ignored editing instructions or collapsed when trying to follow them. Across different types of reference difficulty, tasks involving cross-domain inputs, large-scale or viewpoint gaps, rare concepts, and multilingual inputs are consistently challenging.

Acknowledgements

We thank Minsu Kim and Heiga Zen for their support with this work and for their review of the initial version of the paper. We also appreciate the funding support from Google Japan. MS was supported by JSPS KAKENHI Grant Number JP23H04974.

References

- [1] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos, 2025. 10
- [2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhao-hai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. 8, 1, 6
- [3] Samyadeep Basu, Mehrdad Saber, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods, 2023. 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
- [5] ByteDance. Doubao. <https://www.doubao.com/>, 2025. 3, 5
- [6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 10
- [7] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xincheng Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiuse Gu, Tiankai Hang, Duojuan Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, Yang Li, Zhenxi Li, Zhimin Li, Jiaxin Lin, Linus, Lu-Hao Liu, Shu Liu, Songtao Liu, Yu Liu, Yuhong Liu, Yanxin Long, Fanbin Lu, Qinglin Lu, Yuyan Peng, Yuanbo Peng, Xiang-Yu Shen, Yi-Ping Shi, Jiale Tao, Yang-Dan Tao, Qianhui Tian, Pengfei Wan, Chunyu Wang, Kai Wang, Lei Wang, Linqing Wang, Lucas Wang, Qixun Wang, Weiyang Wang, Hao Wen, Bing Wu, Jianbing Wu, Yue Wu, Senhao Xie, Fangzhou Yang, Miles Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Zhengang Yuan, Chao Zhang, Jianwei Zhang, Pei pei Zhang, Shiyuan Zhang, Tao Zhang, Weigang Zhang, Yepeng Zhang, Yingfang Zhang, Zihao Zhang, Zijian Zhang, Penghao Zhao, Zhiyuan Zhao, Xuefei Zhe, Jian-Xiang Zhu, and Zhao Zhong. Hunyuanimage 3.0 technical report. *ArXiv*, abs/2509.23951, 2025. 2
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3
- [9] Ruoxi Chen, Dongping Chen, Siyuan Wu, Sinan Wang, Shiyun Lang, Peter Sushko, Gaoyang Jiang, Yao Wan, and Ranjay Krishna. Multiref: Controllable image generation with multiple visual references. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13325–13331, 2025. 10
- [10] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 10
- [11] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. *arXiv preprint arXiv:2403.05139*, 2024. 2, 10
- [12] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models, 2024. 10
- [13] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. 2
- [14] Antoine Dussolles, Andrea Cardeña Díaz, Shota Sato, and Peter Devine. M-IFEval: Multilingual instruction-following evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6176–6191, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 10
- [15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yan-nik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2, 10
- [16] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models, 2024. 2, 10
- [17] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024. 6

- [18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [19] Gemini Team. Gemini: A family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 6, 7, 1, 5
- [20] Sara Ghazanfari, Wei-An Lin, Haitong Tian, and Ersin Yumer. Spotedit: Evaluating visually-guided image editing methods. *arXiv preprint arXiv:2508.18159*, 2025. 10
- [21] Google DeepMind. Veo 2, 2024. 10
- [22] Google DeepMind. Nano banana: Gemini 2.5 flash image model. <https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/>, 2025. Accessed: 2025-10-31. 2, 3, 4, 6, 5
- [23] Keno Harada, Yudai Yamazaki, Masachika Taniguchi, Edison Marrese-Taylor, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. When instructions multiply: Measuring and estimating LLM capabilities of multiple instructions following. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16506–16526, Suzhou, China, 2025. Association for Computational Linguistics. 3, 10
- [24] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multiturn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. 3, 10
- [25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 10
- [26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2
- [27] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 10
- [28] Daichi Horita, Naoto Inoue, Kotaro Kikuchi, Kota Yamaguchi, and Kiyoharu Aizawa. Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation. In *CVPR*, 2024. 2, 10
- [29] Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. Videoimage: Multi-subject and motion customization of text-to-video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17603–17612, 2025. 10
- [30] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. 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 (CVPR)*, pages 8362–8371, 2024. 2
- [31] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12469–12478, 2024. 10
- [32] Google Imagen-Team, Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluís Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Rory Lawton, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff, Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dekhtyarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Goyal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Jonathan Heek, Amir Hertz, Ed Hirst, Emiel Hoogeboom, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovan Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun, Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Matthieu Kim Lorrain, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Thomas Mensink, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, Signe Nørly, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Tim Salimans, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Eleni Shaw, Gregory Shaw, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, Evgeny Sluzhaev, Valerii Sokolov, Thibault Sottiaux, Flo-

- rian Stimberg, Brad Stone, David Stutz, Yu-Chuan Su, Eric Tabellion, Shuai Tang, David Tao, Kurt Thomas, Gregory Thornton, Andeep Toor, Cristian Udrescu, Aayush Upadhyay, Cristina Vasconcelos, Alex Vasiloff, Andrey Voynov, Amanda Walker, Luyu Wang, Miaosen Wang, Simon Wang, Stanley Wang, Qifei Wang, Yuxiao Wang, Ágoston Weisz, Olivia Wiles, Chenxia Wu, Xingyu Federico Xu, Andrew Xue, Jianbo Yang, Luo Yu, Mete Yurtoglu, Ali Zand, Han Zhang, Jiageng Zhang, Catherine Zhao, Adilet Zhaxybay, Miao Zhou, Shengqi Zhu, Zhenkai Zhu, Dawn Bloxwich, Mahyar Bordbar, Luis C. Cobo, Eli Collins, Shengyang Dai, Tulsee Doshi, Anca Dragan, Douglas Eck, Demis Hassabis, Sissie Hsiao, Tom Hume, Koray Kavukcuoglu, Helen King, Jack Krawczyk, Yeqing Li, Kathy Meier-Hellstern, Andras Orban, Yury Pinsky, Amar Subramanya, Oriol Vinyals, Ting Yu, and Yori Zwols. Imagen 3, 2024. 10
- [33] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. LayoutDM: Discrete Diffusion Model for Controllable Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10167–10176, 2023. 2, 10
- [34] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. InfiniteYou: Flexible photo recrafting while preserving your identity. In *ICCV*, 2025. 2
- [35] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3, 10
- [36] Jeongho Kim, Guojung Gu, Minhoo Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024. 10
- [37] Akio Kodaira, Chenfeng Xu, Toshiaki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuho, Soichi Sugano, Hanying Cho, Zhijian Liu, Masayoshi Tomizuka, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12371–12380, 2025. 10
- [38] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2023. 6
- [39] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. 10
- [40] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. 3, 10
- [41] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2, 10
- [42] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 5, 10
- [43] LAION-AI. aesthetic-predictor, 2022. 4
- [44] DONGXU LI, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, pages 30146–30166. Curran Associates, Inc., 2023. 2
- [45] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv:2301.07093*, 2023. 10
- [46] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 10
- [47] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022. 2
- [48] Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, Rongrong Ji, et al. I2ebench: A comprehensive benchmark for instruction-based image editing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3
- [49] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2063–2072, 2025. 10
- [50] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 2
- [51] Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, and Jinjia Zhou. Tkg-dm: Training-free chroma key content generation diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13031–13040, 2025. 2, 10
- [52] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 10
- [53] Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your own human image generation model via direct preference optimization with ai feedback. *arXiv preprint arXiv:2405.20216*, 2024. 6

- [54] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 5
- [55] OpenAI. Sora, 2024. 10
- [56] OpenAI. Gpt-4o image generation. <https://openai.com/index/introducing-4o-image-generation/>, 2025. Accessed: 2025-10-31. 2, 3, 4, 6, 5
- [57] OpenAI. Gpt-5. <https://openai.com/ja-JP/index/introducing-gpt-5/>, 2025. Accessed: 2025-11-14. 6, 7, 1
- [58] Y. Oshima, M. Suzuki, Y. Matsuo, and H. Furuta. Inference-time text-to-video alignment with diffusion latent beam search, 2025. *arXiv preprint arXiv:2501.19252*. 6
- [59] 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. 10
- [60] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 10
- [61] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation, 2023. 10
- [62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3, 4
- [63] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 10
- [64] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. 4, 2
- [65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022. 2, 10
- [66] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arxiv:2208.12242*, 2022. 2, 3, 10
- [67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 10
- [68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 4, 2
- [69] 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 (CVPR)*, pages 8871–8879, 2024. 2, 3
- [70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, 2015. 2
- [71] Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. Realedit: Reddit edits as a large-scale empirical dataset for image transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13403–13413, 2025. 10
- [72] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 2
- [73] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 4, 2
- [74] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 10
- [75] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22532–22541, 2023. 10
- [76] Meiyun Wang, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Lost in the distance: Large language models struggle to capture long-distance relational knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4536–4544, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 10
- [77] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18359–18369, 2023. 2, 3

- [78] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. [2](#), [10](#)
- [79] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. [2](#), [6](#), [5](#)
- [80] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yuezhang Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. [2](#), [3](#), [6](#), [5](#)
- [81] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. [2](#)
- [82] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Boosting text-to-video generative model with MLLMs feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [6](#)
- [83] Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni2: Multimodal instruction-based editing and generation, 2025. [2](#), [3](#), [4](#), [6](#), [5](#)
- [84] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing, 2025.
- [85] Shitao Xiao, Yuezhang Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. [2](#)
- [86] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Oot-diffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. [10](#)
- [87] Zhipeng Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, 2025. [3](#)
- [88] Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Image referenced sketch colorization based on animation creation workflow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23391–23400, 2025. [10](#)
- [89] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. [10](#)
- [90] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapt: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. [2](#), [10](#)
- [91] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. [2](#), [3](#), [4](#), [6](#), [5](#), [9](#)
- [92] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26125–26135, 2025. [3](#), [5](#)
- [93] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023. [2](#), [3](#), [5](#)
- [94] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. [10](#)
- [95] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. [10](#)
- [96] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. [10](#)
- [97] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. [3](#), [10](#)
- [98] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4606–4615, 2023. [2](#), [10](#)