Generative Multimodal Pretraining with Discrete Diffusion Timestep Tokens

Supplementary Material

Overview In this supplementary material, we present the following content:

A. Implementation Details

A.1. Tokenizer

Encoder. In the encoder, we set $T = 480$ (*i.e.*, the number of query tokens). The dimensions of the query tokens and the noise-free image are $R^{480\times256}$ and $R^{1024\times256}$, respectively. The encoder contains two independent transformers, each comprising 20 layers with latent dimension of 256. Following SD3 [14], despite the noise-free images and query tokens being input into separate transformers, we join the sequences of the two for the attention operation. This allows both representations to operate independently while considering the influence of the other. The encoder output retains only the transformed query tokens, serving as the image's latent representations.

Quantizer. The quantizer is an EMA-variant of vector quantization. Following [64], we leverage a linear projection from the encoder output to low-dimensional variable space for code index lookup (*i.e.*, reduced from a 256-d vector to a 16-d vector per code). We also apply L2 normalization on the encoded latent features and codebook latent variables. Moreover, at each training step, we reset the dead entries in the codebook $\mathcal{C}$ (*i.e.*, rarely matched with any tokens) to random tokens in the training batch.

Table 1. The detailed training hyper-parameters. "MLLM-pt" denotes the pretraining of DDT-LLaMA, "MLLM-ft" denotes the instruction tuning of DDT-LLaMA, while "Tokenizer" denotes the training of DDT.

Hyper-parameters	MLLM-pt	MLLM-ft	Tokenizer
LLM init	LLama3-8B	MLLM-pt	-
Optimizer	AdamW	AdamW	AdamW
Optimizer param.	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1\text{e}{-6}$		$\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 1\text{e}{-6}$
Peak LR	1e-4	1e-5	1e-4
LR scheduler	Cosine	Cosine	Linear+Cosine
Batch size	1280	256/128	1024
Training Steps	360K	160K	140K
Warmup Steps	5K	2K	5K
Weight decay	0.05	0.05	0.0
Gradient clipping	1.0	1.0	-
Numerical precision	bfloat16	bfloat16	bfloat16
Resource Usage	512 Ascend 910B	256 Ascend 910B	32 NVIDIA A800
Framework	Megatron(TP=8)	Megatron(TP=8)	DDP

Decoder. we use the same MMDiT architecture proposed in SD3 [14] for our decoder with minor modifications. Each transformer in the MMDiT comprises 24 layers with latent dimension of 1536. The sequence of quantized tokens replaces the text tokens as input, with a linear layer to project the 16-dimensional quantized vector to the latent dimension (*i.e.*, 1536) of the MMDiT. Additionally, we also removed the pooled token embedding introduced in SD3.

Furthermore, the hyper-parameters of training the tokenizer are detailed in **Table 1**.

A.2. MLLM

We initialize our MLLM from a pretrained LLM, specifically using the Llama3-8b model [12], which has only undergone pretraining *without instruction tuning*. Additionally, we expand its vocabulary by adding $|\mathcal{C}| = 65,536$ visual codes and two extra special tokens (`[BOV]` and `[EOV]`). Since both image and text are represented as discrete token IDs, we can use the cross-entropy to supervise the token prediction at each position for both modalities with a shared prediction head. A shared head has proven more effective than using separate heads for each modality in training, enhancing the upper capabilities of the MLLM. During inference, when generating content for a specific modality, tokens that do not fall in that modality's space should be masked. For example, when generating text, we mask the logits of the $65,536$ visual codes before sampling the text tokens; Similarly, when the `[BOV]` token indicates the beginning of image output, we mask the logits for $128,256$ text words before sampling the visual tokens. Moreover, we set $topk = 50, topp = 1.0$ for text token sampling and $topk = 4096, topp = 0.9$ for visual token sampling. Besides, for text-to-image generation, during inference we use classifier-free guidance on the logits for autoregressive sampling in a manner similar to [42, 59]. We

Figure 1. More qualitative results of DDT-LLaMA text-to-image generation. (*the supplement to Figure 3 in the main paper*)

set the guidance scale to 8.0.

During both pretraining and instruction-tuning, all parameters of the MLLM are **fully fine-tuned**. The hyperparameters of both training stages for DDT-LLaMA are shown in **Table 1**. The overall training is stable; we observed only one minor spike in the loss curve during pretraining. Following [9], we resume training from a checkpoint approximately 500 steps before the onset of the spike.

B. Training Data & Evaluation Details

B.1. Training Data

Tokenizer Training. Our diffusion timestep tokenizer is trained on the training split of ImageNet [11], which comprises about 1.28 million images. Besides, each training image is center-cropped to a size of 256×256. **The training was conducted on 32 NVIDIA A800 GPUs and lasted for nearly one weak.**

MLLM Pretraining. Our pretraining dataset, sourced from Laion [54], consists of 200 million text-to-image pairs. Each pair includes an image accompanied by a brief, coarse-grained native caption and a detailed, fine-grained generated caption. For the generation of detailed captions, we employ ShareGPT-4V [8] to annotate over 400 million images sourced from Laion [54] and Coyo [**?**]. Considering the propensity of ShareGPT-4V to generate captions that may not align with the images due to hallucinations, we utilize CLIP scores [50] to filter all generated image-text pairs, retaining only those with the highest CLIP scores, totaling 200 million images. For pretraining, we structure each pair in the format: "`[BOS] <caption text> [BOV] <DDT tokens> [EOV] [EOS]`" for pretraining, where `[BOS]` and `[EOS]` are the original special tokens from the text tokenizer, `[BOV]` and `[EOV]` marking the start and the end of the vision input. Each image has a 60% chance of being paired with the long caption and a 40% probability of being paired with the short caption during training. Besides, each sample's caption has a 10% probability of being dropped out. Furthermore, to preserve the textual capabilities of MLLM, we supplement our dataset with purely textual data from Wikipedia and Pile [17], at a ratio of 10%. **The pre-training was conducted on 512 Ascend 910B NPUs and lasted for nearly two weaks.**

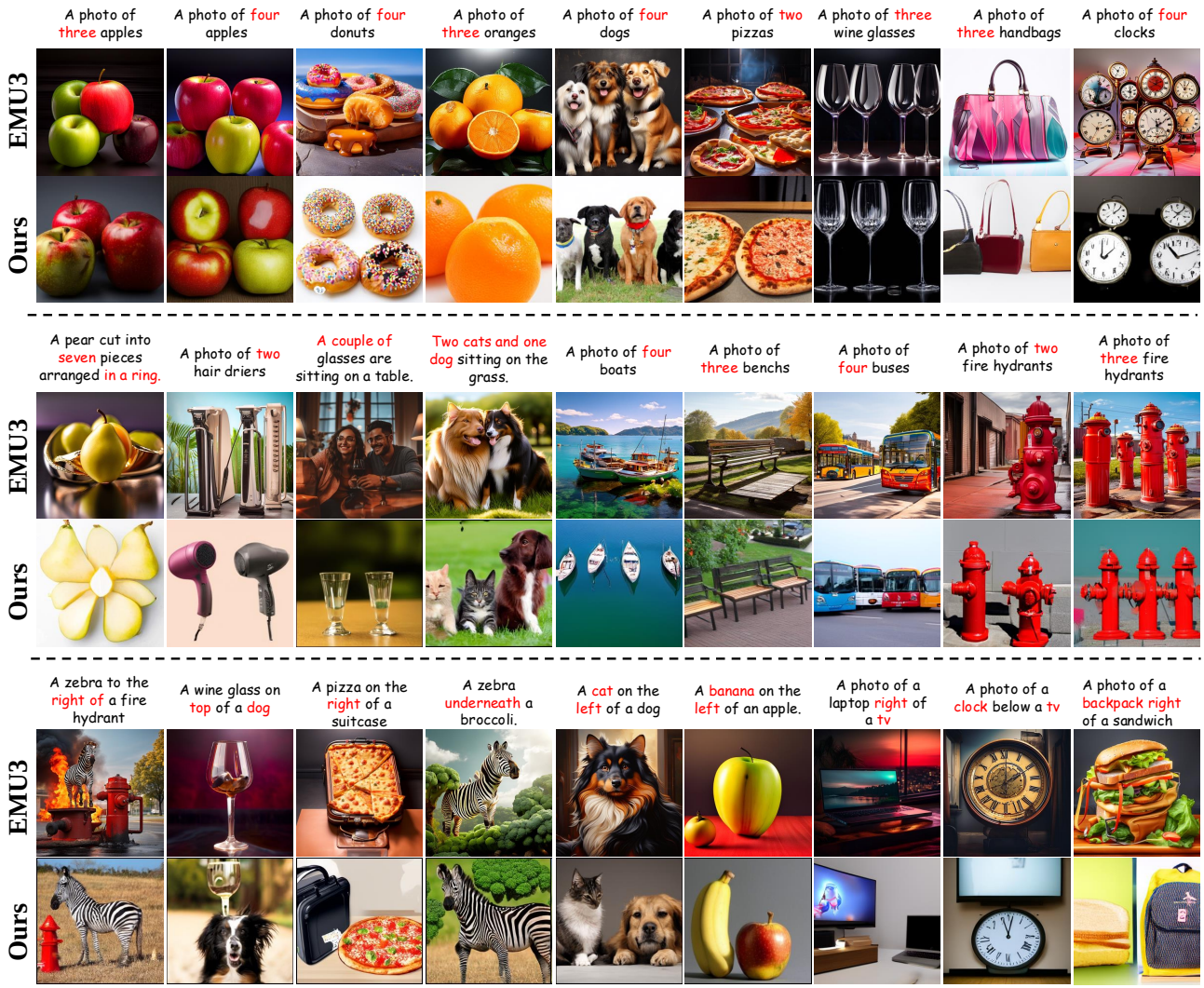

Figure 2. More qualitative comparison with EMU3 on T2I generation (PART-1). DDT-LLaMA can better respond to prompts related to counting, color, and position. (*the supplement to Figure 4 in the main paper*)

MLLM Instruction Tuning. During instruction tuning, we incorporate a variety of tasks, outlined as follows: (1) Text-to-Image Generation: We employ datasets including ShareGPT4V caption [8], ALLaVA [6] and GRIT [6], utilizing a prompt template formatted as:"`[BOS] USER: <caption> Please Generate an image. ASSISTANT: [BOV] <DDT tokens> [EOV] [EOS]`".

(2) Image editing: We employ datasets such as InstructPix2Pix [5] and Hive [70], utilizing a prompt template formatted as:"`[BOS] USER: [BOV] <input DDT tokens> [EOV] <instruction> Please Generate an image. ASSISTANT: [BOV] <output DDT tokens> [EOV] [EOS]`". Besides, considering the generally mediocre quality of existing image editing datasets, we also construct a batch of higher quality image editing data for integration into the training.

(3) Image caption & VQA: We mainly leverage ShareGPT4V(-instruct) [8], ALLaVA(-instruct) [6] and select part of the held-in training datasets in InstructBlip [10] for instruction-tuning. We follow the [10] to design the instruction templates.

B.2. Evaluation Details

Baseline Methods. For text-to-image generation tasks, we compare DDT-LLaMA with both diffusion-based T2I specialists and MLLM-based generalists. The diffusion-based T2I specialists include DALL-E 2 [51], SDv1.5 [52], SDv2.1 [52], SDXL [48], PixArt-alpha [7], DALL-E 3 [3], and SD3 [14]. For the MLLM-based generalists, we include comparisons with SEED-LLaMA [18], LaVIT [27], Emu2-Gen [56], SEED-X [19], VILA-U [62], Lumina-mGPT [39], and Emu3 [59]. ***(For EMU3, we report its results without prompt rewriting for fair comparison.)***

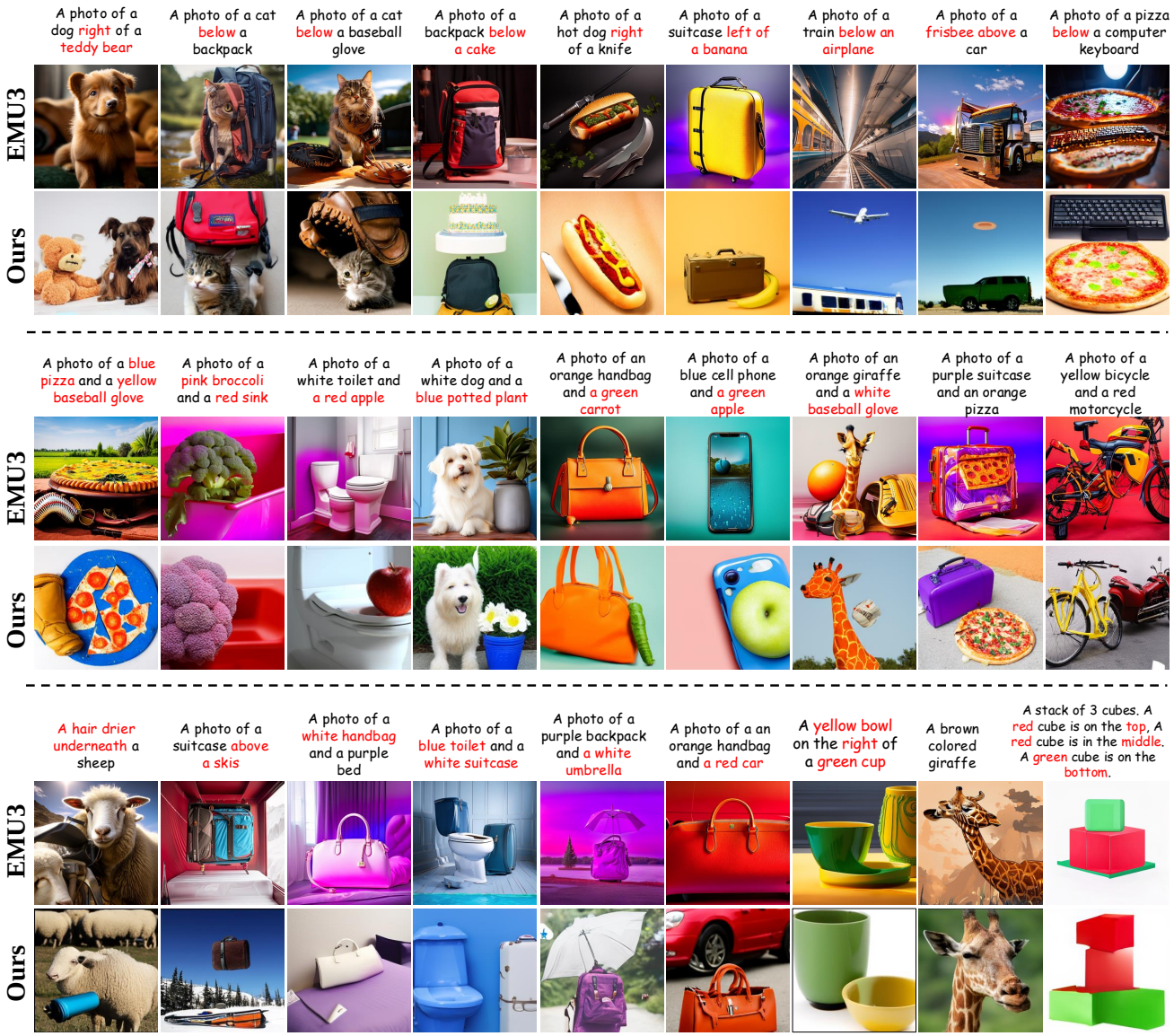

Figure 3. More qualitative comparison with EMU3 on T2I generation (PART-2). DDT-LLaMA can better respond to prompts related to counting, color, and position. (*the supplement to Figure 4 in the main paper*)

For image editing tasks, we compare DDT-LLaMA with both specialized image editing models and generalist MLLM-based models. The image editing specialists we evaluate include InsPix2Pix [5], MGIE [16], and UltraEdit [71]. Among the MLLM-based generalist models, we compare against GILL [28], Emu2-Gen [56], SEED-LLLaMA [18], LaVIT [27], and SEED-X-Edit [19]. ***(We exclude MLLMs such as EMU3 because they lack image editing capabilities.)***

For visual comprehension and generation tasks, we compare DDT-LLaMA with specialized visual comprehension MLLMs, and MLLMs capable of both visual comprehension and generation. The specialized visual comprehension models include InstructBlip[10], QWenVL-Chat [2], LLaVA-1.5 [40], mPLUG-Owl2 [63], ShareGPT4V [8], LLaVA-1.6(HD) [41], and VILA [34]. For models supporting both visual comprehension and generation, we compare DDT-LLaMA against Emu2-Chat [56], SEED-LLLaMA [18], VILA-U [62], LaVIT [27], and Emu3 [59].

Evaluation Dataset. For text-to-image generation tasks, we conduct zero-shot evaluation on 3 benchmarks: GenEval [20], T2I-CompBench [24], and DrawBench [53]. GenEval contains 6 different subtasks of varying difficulty requiring various compositional skills, including `single object` (SingObj), `single object` (TwoObj), `counting`, `colors`, `position`, `color`

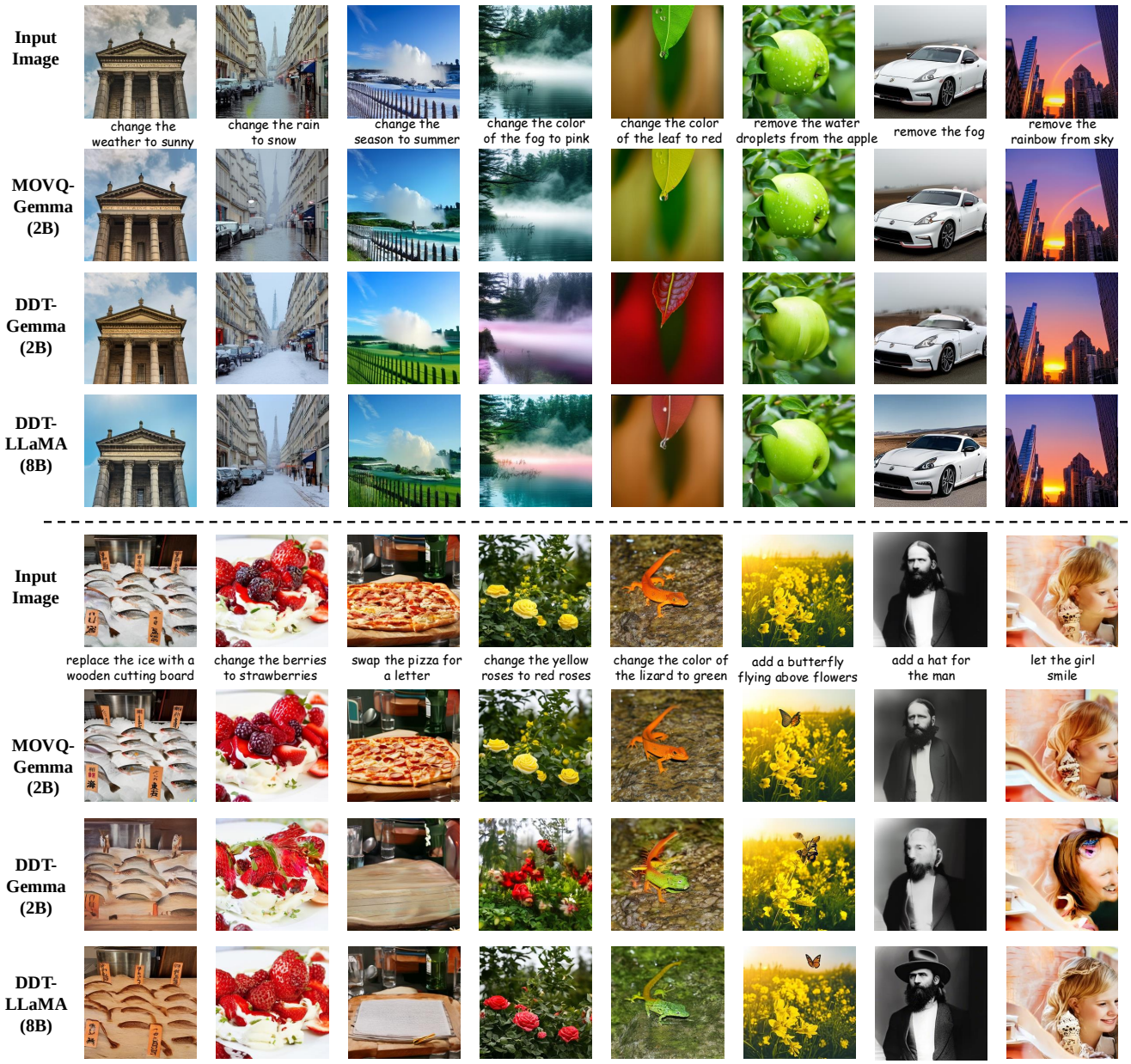

Figure 4. Qualitative comparison of MOVQ-Gemma(2B), DDT-Gemma(2B), and DDT-LLaMA(8B) in the image editing task. In most instances, DDT-Gemma outperforms MOVQ-Gemma. Furthermore, DDT-LLaMA not only effectively comprehends and executes editing instructions accurately but also excels in preserving image fidelity.

`binding` (ColorAttri). And we adopt the metric proposed by [20] for evaluation. Each subtask is scored independently, and the overall score is calculated as the average of all six subtask scores. The T2I-CompBench suite encompasses six subtasks: `color`, `shape`, `texture`, `spatial`, `non-spatial`, and `complex` (complex compositions). Building on prior research, we employ the Blip-VQA score [30] to assess the color, shape, and texture subtasks. For spatial evaluation, we use the UniDet score [73]; for non-spatial evaluations, the CLIP score [23, 50]; and for complex compositions, the 3-in-1 Metric [24]. In terms of DrawBench, we leverage Clip text-visual feature similarity [50] as the evaluation metric.

We also conduct zero-shot instruction-based image editing across three datasets: EVR [57], MA5k [55], and MagicBrush [68]. Following [16], for EVR and MagicBrush, we treat the standard pixel difference (L1) and visual feature similarity from the CLIP visual encoder (CVS) between generated images and ground-truth goals as the evaluation metrics. For MA5K, we utilize L1 and LPIPS [69] as the

Figure 5. (More examples of text-to-image generation with different MLLM size (2B, 8B) and training compute (50%, 75%, 100% of total tokens). *the supplement to Figure 7 in Section 5.4.5 of the main paper*)

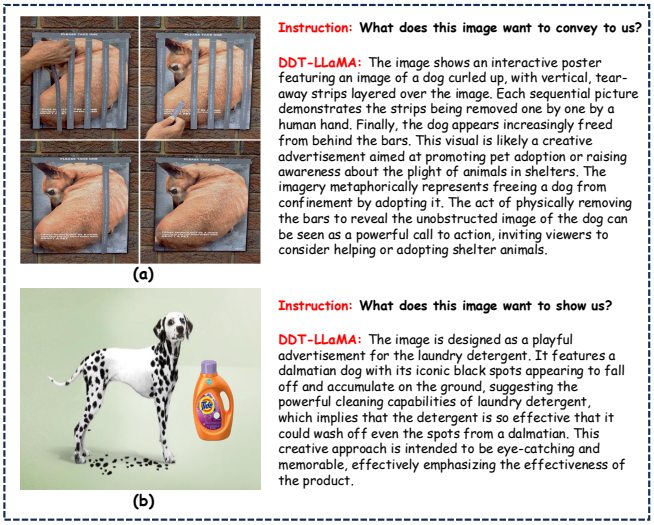

Figure 6. Qualitative results of DDT-LLaMA visual comprehension.

evaluation metrics.

For visual comprehension tasks, we conduct zero-shot evaluation on a wide range of academic benchmarks, including image caption (NoCaps [1], Flickr30K [47]), VQA (VQAv2 [21], GQA [26], OKVQA [43], VizWiz [4]), MLLM-oriented Comprehension Benchmarks (MME [15], SEEDBench [29], POPE [33]). (Note: In Table3 of the main paper, we use "VQA" to denote VQAv2). We employ CIDEr as the metric for image caption tasks, and VQA accuracy for VQA tasks. We employ the CIDEr metric to evaluate performance on image captioning tasks, while using VQA accuracy for the VQA datasets. Moreover, each MLLM-oriented benchmark is evaluated according to its specific prescribed methodologies, where we report the perception score for MME, MCQ accuracy for SEEDBench, and the F1 score for POPE.

C. Additional Examples of T2I Generation

In **Figure 1**, we present more qualitative examples of DDT-LLaMA on text-to-image generation tasks. DDT-LLaMA adeptly handles various types of instructions, including complex ones such as generating surreal images (*e.g.*, "A panda drinking coffee", " sloth with pink hat") and multi-

Figure 7. More results of counterfactual interpolation with DDT tokens and VQGAN tokens. (*the supplement to Figure 6 in Section 5.4.2 of the main paper*)

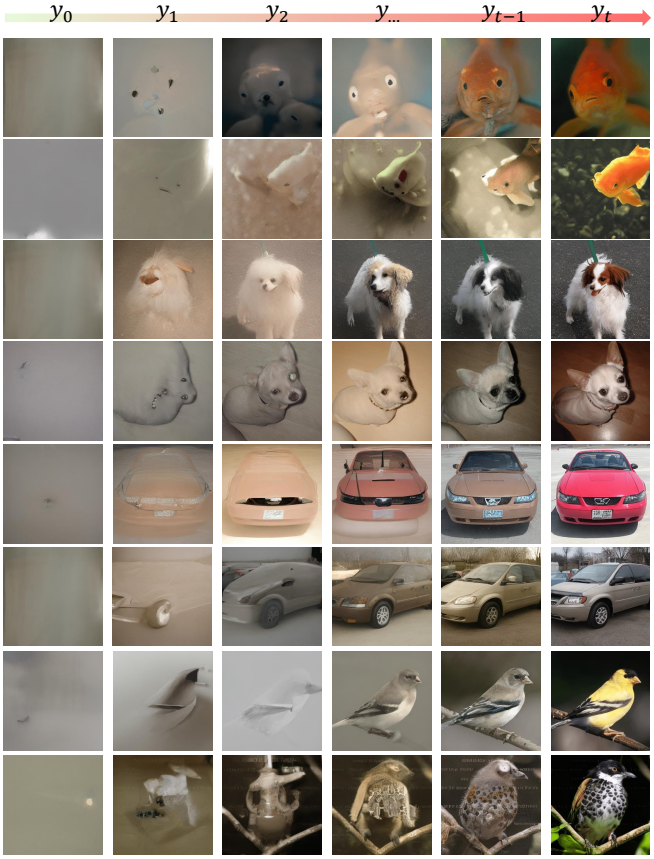

Figure 8. More results of decoding images with an expanding subset of autoregressive-sampled DDT tokens (from 1 to $T = 480$). y_k denotes the number of sampled DDT tokens that are fed into the decoder for image generation, $1 = y_0 < y_1 < y_2 < ... < y_{t-1} < y_t = 480$. (*the supplement to Figure 8 in Section 5.4.2 of the main paper*)

condition combined prompts (*e.g.*, "An emoji of a baby panda wearing a red hat, blue gloves, green shirt, and blue pants"), to generate semantically-consistent images.

Furthermore, in **Figure 2** and **Figure 3**, we present a direct comparison between DDT-LLaMA and Emu3 across 54 prompts involving counting, color, and positioning. It is evident that Emu3 falls short in these areas: (1) For counting-related prompts, EMU3 often generates images with *an incorrect number of objects*. (2) For prompts related to positioning, Emu3 frequently *misplaces objects*, and sometimes *only one of the objects* is generated. (3) For color-related prompts, EMU3 often *incorrectly assigns colors to the objects*, and it may also generate images where the arrangement or presence of objects is *disordered* ("A photo of a purple suitcase and an orange pizza" in Figure 3). In contrast, DDT-LLaMA generates images that **more accurately reflect the desired object attributes (number and color) and adhere to the spatial specifications outlined in the prompts**.

D. Additional Comparison on Image Editing

As discussed in Section 5.4.4, we also employ Gemma2-2b [58] as the initial LLM and leverage both DDT tokens and MoVQ tokens for pretraining and instruction tuning, which we refer to as DDT-Gemma and MOVQ-Gemma, respectively. In **Figure 4**, we showcase a series of qualitative examples that compare MOVQ-Gemma (2B), DDT-Gemma (2B), and DDT-LLaMA (8B). First, when comparing MOVQ-Gemma and DDT-Gemma, it is evident that in many editing cases, *MOVQ-Gemma often gives up editing*, typically returning the original image as the output. In contrast, DDT-Gemma exhibits a more robust comprehension of the editing instructions and delivers superior results in

A/B tests, which indicates that **our recursive DDT tokens outperform spatial tokens in image editing tasks**.

Furthermore, DDT-Gemma sometimes faces problems with incomplete modifications or fails to maintain image fidelity in areas not targeted by the edits. For example, in the 7th case of Figure 4 ("remove the fog"), only the red exhaust behind the car is eliminated by DDT-Gemma. In the 12th case ("change the yellow roses to red roses"), the shape of the roses is also inadvertently changed by DDT-Gemma. In contrast,**scaling up the backbone model from 2B to 8B significantly improves the editing performance**. We can see that DDT-LLaMA not only effectively comprehends the instructions to **accurately execute the editing**, but also excels at **preserving image fidelity**. This serves as evidence of the scaling-law properties of DDT tokens.

E. Examples of Visual Comprehension

In **Figure 6**, we show some qualitative examples of visual comprehension. Although lack of pretrained encoders like CLIP [50], DDT-LLaMA can still effectively understand the visual semantics in the images and accurately infer the answer based on the given textual instruction.

F. Additional Examples of In-depth Analysis

F.1. Counterfactual Interpolation

Figure 7 shows more results of counterfactual interpolation of VQGAN tokens [13] and DDT tokens. DDT tokens employ a disentangled representation to ensure that only the attributes represented by the substituted tokens vary in the generated counterfactuals, which **allows for a seamless semantic integration of the two images**.

F.2. Decoding with a subset of DDT tokens

In **Figure 8**, we show more results demonstrating how autoregressive-sampled DDT tokens can be decoded into images in order. As the number of sampled tokens increases (from 1 to $T = 480$), the image attributes are progressively reconstructed – from fine details to the completion of coarse-grained contours and color information. This confirms that **our DDT token sequence successfully decouples image attributes and possesses recursive properties**.

F.3. Scaling Laws of DDT-based MLLM

In **Figure 5**, we show more examples of text-to-image generation examples using two model sizes (Gemma 2B, LLama 8B [12]) at three different training stages (50%, 75%, and 100% of total training tokens). The enhancements in visual quality observed correspond with scaling laws, which indicate that **larger transformers trained on more comprehensive datasets tend to yield superior text-to-image performance**.

G. Limitation and Future Work

Our current tokenizer is trained solely on the ImageNet dataset [11] at a resolution of 256x256 pixels, and it faces limitations in reconstructing open-domain images compared to baseline methods. For example, the EMU3 tokenizer MOVQ [72], which is trained on a significantly larger dataset (from Laion and InternVID [60]), achieves superior reconstruction performance than ours. As our 200M pretraining dataset filtered from an open-domain image dataset (*i.e.*, Laion), the inadequate reconstructive capability of DDT restricts DDT-LLaMA's ability in text-to-image generation. This particularly impacts the aesthetic quality of the images generated by DDT-LLaMA, as illustrated in Figure 1, Figure 2, and Figure 3.

We are currently working on improving and scaling up the training of our DDT-tokenizer and the MLLM on a significantly larger dataset (about 500M images). ***In the near future, we will release a more powerful version of DDT-LLaMA, along with a detailed technical report. Stay tuned!*** Building on this foundation, we aim to further demonstrate that DDT-LLaMA is a significant approach for addressing visual-language tasks [31, 45, 65–67] and collaborative NLP tasks [22, 44, 46, 61, 74, 75]. We also seek to extend the capabilities of DDT-LLaMA to support more vision-language tasks [25, 32, 35–38, 49] such as video comprehension and video generation.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 6

[2] 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*, 2023. 4

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 6

[5] 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. 3, 4

[6] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 3

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 3

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2, 3, 4

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 2

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3, 4

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 8

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 8

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 8

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 3

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 6

[16] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 4, 5

[17] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 2

[18] 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. 3, 4

[19] 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. 3, 4

[20] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 5

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 6

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. 8

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 5

[24] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 4, 5

[25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. 8

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6

[27] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Yadong Mu, et al. Unified language-vision pre-training in llm with dynamic discrete visual tokenization. In *International Conference on Learning Representations*, 2024. 3, 4

[28] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[29] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5

[31] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to

follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 2023. 8

[32] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu, and Yueting Zhuang. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12601–12617, 2023. 8

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 6

[34] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 4

[35] Wang Lin, Jingyuan Chen, Jiaxin Shi, Zirun Guo, Yichen Zhu, Zehan Wang, Tao Jin, Zhou Zhao, Fei Wu, YAN Shuicheng, et al. Action imitation in common action space for customized action image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 8

[36] Wang Lin, Tao Jin, Wenwen Pan, Linjun Li, Xize Cheng, Ye Wang, and Zhou Zhao. Tavt: Towards transferable audio-visual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14983–14999, 2023.

[37] Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao. Exploring group video captioning with efficient relational approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15281–15290, 2023.

[38] Wang Lin, Jingyuan Chen, Jiaxin Shi, Yichen Zhu, Chen Liang, Junzhong Miao, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng Yan, et al. Non-confusing generation of customized concepts in diffusion models. *arXiv preprint arXiv:2405.06914*, 2024. 8

[39] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 3

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 4

[41] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4

[42] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pages arXiv–2402, 2024. 1

[43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 6

[44] Kaihang Pan, Juncheng Li, Hongye Song, Jun Lin, Xiaozhong Liu, and Siliang Tang. Self-supervised meta-prompt learning with meta-gradient regularization for few-shot generalization. *arXiv preprint arXiv:2303.12314*, 2023. 8

[45] Kaihang Pan, Zhaoyu Fan, Juncheng Li, Qifan Yu, Hao Fei, Siliang Tang, Richang Hong, Hanwang Zhang, and Qianru Sun. Towards unified multimodal editing with enhanced knowledge collaboration. *Advances in Neural Information Processing Systems*, 37:110290–110314, 2024. 8

[46] Kaihang Pan, Juncheng Li, Wenjie Wang, Hao Fei, Hongye Song, Wei Ji, Jun Lin, Xiaozhong Liu, Tat-Seng Chua, and Siliang Tang. I3: I ntent-i ntrospective retrieval conditioned on i nstructions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1839–1849, 2024. 8

[47] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[48] 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. *arXiv preprint arXiv:2307.01952*, 2023. 3

[49] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms' compositional reasoning by spatio-temporal graph-guided self-training. *arXiv preprint arXiv:2412.00161*, 2024. 8

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 8

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 4

[54] 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. 2

[55] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning:

Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 5

[56] 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. 3, 4

[57] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 5

[58] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 7

[59] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1, 3, 4

[60] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. 8

[61] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. 8

[62] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 3, 4

[63] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 4

[64] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 1

[65] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21560–21571, 2023. 8

[66] 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. *arXiv preprint arXiv:2411.15738*, 2024.

[67] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12944–12953, 2024. 8

[68] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023. 5

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[70] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9026–9036, 2024. 3

[71] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie 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. 4

[72] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022. 8

[73] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection, 2022. 5

[74] Yun Zhu, Jianhao Guo, Fei Wu, and Siliang Tang. Rosa: A robust self-aligned framework for node-node graph contrastive learning. *arXiv preprint arXiv:2204.13846*, 2022. 8

[75] Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. *arXiv preprint arXiv:2410.10329*, 2024. 8