Planting a SEED of Vision in Large Language Model

Yuying Ge[1⋆] **Yixiao Ge**[1,2⋆†] **Ziyun Zeng**[2] **Xintao Wang**[1,2] **Ying Shan**[1,2]

[1]Tencent AI Lab [2]ARC Lab, Tencent PCG

https://github.com/AILab-CVC/SEED

Abstract

We present **SEED**, an elaborate image tokenizer that empowers Large Language Models (LLMs) with the emergent ability to **SEE** and **D**raw at the same time. Research on image tokenizers has previously reached an impasse, as frameworks employing quantized visual tokens have lost prominence due to subpar performance and convergence in multimodal comprehension (compared to BLIP-2, etc.) or generation (compared to Stable Diffusion, etc.). Despite the limitations, we remain confident in its natural capacity to unify visual and textual representations, facilitating scalable multimodal training with LLM's original recipe. In this study, we identify two crucial principles for the architecture and training of SEED that effectively ease subsequent alignment with LLMs. (1) Image tokens should be independent of 2D physical patch positions and instead be produced with a *1D causal dependency*, exhibiting intrinsic interdependence that aligns with the left-to-right autoregressive prediction mechanism in LLMs. (2) Image tokens should capture *high-level semantics* consistent with the degree of semantic abstraction in words, and be optimized for both discriminativeness and reconstruction during the tokenizer training phase. As a result, the off-the-shelf LLM is able to perform both image-to-text and text-to-image generation by incorporating our SEED through efficient LoRA tuning. Comprehensive multimodal pretraining and instruction tuning, which may yield improved results, are reserved for future investigation. This version of SEED was trained in 5.7 days using only 64 V100 GPUs and 5M publicly available image-text pairs. Our preliminary study emphasizes the great potential of discrete visual tokens in versatile multimodal LLMs and the importance of proper image tokenizers in broader research.

1 Introduction

In recent years, Large Language Models [1, 2, 3] (LLMs) pre-trained on massive text corpus with straightforward training objectives such as next-word prediction have exhibited remarkable abilities to understand, reason, and generate texts across a variety of open-ended tasks. Recent studies further exploit the strong generality of LLMs to improve visual understanding or generation tasks, collectively referred to as Multimodal LLM (MLLM). For example, previous work [4, 5, 6, 7, 8] perform open-ended visual QAs through aligning visual features of a pre-trained image encoder (e.g., CLIP-ViT) with the input embedding space of LLMs. GILL [9] empowers LLM with the image generation ability by aligning its output embedding space with the pre-trained Stable Diffusion (SD) model [10].

While these studies have contributed to technological advancements, MLLMs have yet to achieve the remarkable success of LLMs in terms of emergent capabilities. We have made a bold assumption that the premise for the emergence of multimodal capabilities is that text and images can be represented

⋆Equal Contribution.
†Correspondence to yixiaoge@tencent.com.

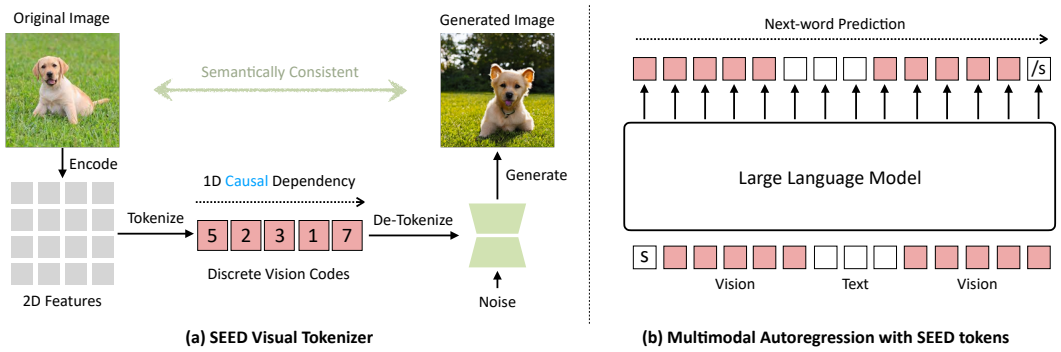

Figure 1: (a) The proposed SEED is a discrete image tokenizer, producing quantized visual codes with 1D causal dependency and high-level semantics. (b) SEED visual tokens enable LLMs to perform both visual comprehension and generation through multimodal autoregression with interleaved image-text data.

and processed **interchangeably** in a unified autoregressive Transformer. Fortunately, we have just found consensus in concurrent works [11, 12], all employing image-to-text and text-to-image generation tasks to demonstrate the emergent ability of unifying visual comprehension and generation in one framework. Regardless of discrete or continuous visual tokens, the training paradigm can be summarised into three stages: **visual tokenizer training, multimodal pretraining, and multimodal instruction tuning**. While concurrent studies primarily emphasize multimodal training (the latter two stages), this work focuses more on the visual tokenizer (the first stage).

We posit that a proper visual tokenizer can facilitate the follow-up multimodal training by (i) easing the semantic alignment between visual and word tokens, and (ii) enabling LLM's original training recipe (i.e., next-word prediction) for multimodal data without specific adaptation for visual tokens. Representing images as a sequence of discrete IDs is naturally compatible with the autoregressive training objective of LLMs. But unfortunately, works [13, 14] that utilize discretized visual tokens for multimodal tasks have receded from prominence, as such models generally rely on super-scale training to converge, leading to substantial training costs. Moreover, we empirically found that the dominant tokenizer VQ-VAE [15] in existing works captures too low-level information for LLMs to effectively perform multimodal comprehension tasks. Existing image tokenizers fail to meet the requirements of unifying visual understanding/generation tasks and facilitating multimodal training.

To this end, we introduce **SEED**, a VQ-based image tokenizer that produces discrete visual codes with 1D causal dependency and necessary high-level semantics for both visual comprehension and generation tasks, as shown in Fig. 1. The off-the-shelf LLMs can be readily equipped with SEED by treating discrete visual tokens as new words and updating the vocabulary with mapped visual codes. In the paper, we present an MLLM by tuning the pre-trained LLM with low-rank adaptation (LoRA) to efficiently align with the SEED tokenizer.

We would like to emphasize the design principles of SEED. (1) *Why causal-dependent tokens?* Existing visual tokens (e.g., from VQ-VAE or CLIP-ViT) are generated using 2D context, which is incompatible with the unidirectional attention in dominant LLMs and counterintuitive for text-to-image tasks requiring raster order prediction. Thus, we convert 2D raster-ordered embeddings into a sequence of semantic codes with 1D causal dependency. (2) *Why high-level semantics?* Since visual and textual tokens in LLMs are expected to be interoperable—sharing weights and training objectives—they should encompass the same degree of semantics to prevent misalignment, i.e., the high-level semantics inherently present in words.*

Specifically, the SEED tokenizer is composed of a ViT encoder, Causal Q-Former, VQ Codebook, Reverse Q-Former, and a UNet decoder. The ViT encoder and UNet decoder are directly derived from the pre-trained BLIP-2 and SD models, respectively. (1) *Tokenize:* Causal Q-Former converts 2D raster-ordered features produced by the ViT encoder into a sequence of causal semantic embeddings,

*While focusing on high-level semantics during tokenization, it is still possible to achieve accurate spatial structural control, such as layout and mask conditions, in image generation tasks. These spatial structural prompts can be tokenized similarly, as demonstrated by the success of SD [10, 16].

which are further discretized by the VQ Codebook. (2) *De-Tokenize:* The discrete visual codes are decoded into generation embeddings via Reverse Q-Former. The generation embeddings are aligned with the latent space of SD so that realistic images with consistent semantics can be generated using the off-the-shelf SD-UNet.

During SEED training, only Causal Q-Former, VQ Codebook, and Reverse Q-Former are tunable. Causal Q-Former is optimized by image-text contrastive loss. VQ Codebook and Reverse Q-Former are trained toward the objectives of dual reconstruction, i.e., the reconstruction between continuous causal embeddings and discrete causal codes, the reconstruction between generation embeddings and the paired textual features. The training objectives ensure that SEED encapsulates the essential semantics for both visual comprehension and generation. Quantitative results indicate that discrete SEED tokens exhibit competitive performance in text-image retrieval compared to BLIP-2, and in image generation compared to Stable Diffusion. With further multimodal autoregressive training, SEED-OPT$_{2.7\mathrm{B}}$ (efficiently tuned via LoRA using 5M image-text pairs) effectively performs image-to-text and text-to-image tasks, yielding promising results in zero-shot image captioning and visual QA, as well as generating high-quality images.

This effort aims to integrate multimodal comprehension and generation tasks within an LLM using discrete visual tokens. Our initial exploration of proper tokenizer designs strives to promote the development of emergent multimodal capabilities. Future work can further scale up training for a better tokenizer and leverage stronger LLMs (e.g., LLaMA [1]) for comprehensive multimodal pretraining and instruction tuning.

2 SEED Visual Tokenizer

2.1 Pilot Experiments of Baseline Tokenizers

Visual tokenizer aims to represent the image as a sequence of discrete tokens. Previous work [15, 13, 17] trains a Vector Quantized Variational AutoEncoders (VQ-VAE) by reconstructing image pixels, while Beit v2 [18] propose vector-quantized knowledge distillation (VQ-KD) to train a visual tokenizer by reconstructing high-level features from the teacher model. We conduct two experiments to respectively align discrete representations of VQ-VAE and Beit v2 with OPT$_{2.7\mathrm{B}}$ [19] model on CC3M [20] dataset. We evaluate the performance with zero-shot image captioning on COCO [21]. VQ-VAE achieves CIDEr 34.0 while Beit v2 achieves 42.0. The experiment results demonstrate that a high-level visual tokenizer, which captures semantic representations of images instead of low-level image details is more effective for multimodal comprehension.

2.2 Architecture

In this work, we introduce a VQ-based image tokenizer **SEED** to produce discrete visual codes with 1D causal dependency and high-level semantics. Specifically, as shown in Fig. 2, the SEED tokenizer is composed of a ViT image encoder [22], Causal Q-Former, VQ Codebook, Reverse Q-Former, and a UNet decoder [10]. The ViT encoder and UNet decoder are directly derived from the pre-trained BLIP-2 and SD models, respectively. We first train a Causal Q-Former to convert 2D raster-ordered features (16×16 tokens) produced by the ViT encoder into a sequence of causal semantic embeddings (32 tokens). We then train a visual codebook to discretize the causal embeddings to quantized visual codes (32 tokens) with causal dependency. We employ a Reverse Q-Former to decode the visual codes into generation embeddings (77 tokens), which are aligned with the latent space of the pre-trained Stable Diffusion (SD) model.

2.2.1 Training Stage I: Causal Q-Former

As shown in Fig. 2, a set number of learnable query embeddings (32 tokens) and features of a pre-trained ViT image encoder are fed into the Causal Q-former to encode a fixed number of causal embeddings (32 tokens) of the input image. Specifically, the query embeddings can interact with only previous queries through self-attention layers with causal mask, and interact with frozen image features through cross-attention layers. We adopt contrastive learning to optimize Causal Q-former fine-tuned from pre-trained BLIP-2 Q-Former on 5M image-text pairs including CC3M [20], Unsplash [23], and COCO dataset [21]. We use contrastive loss to maximize the similarity between

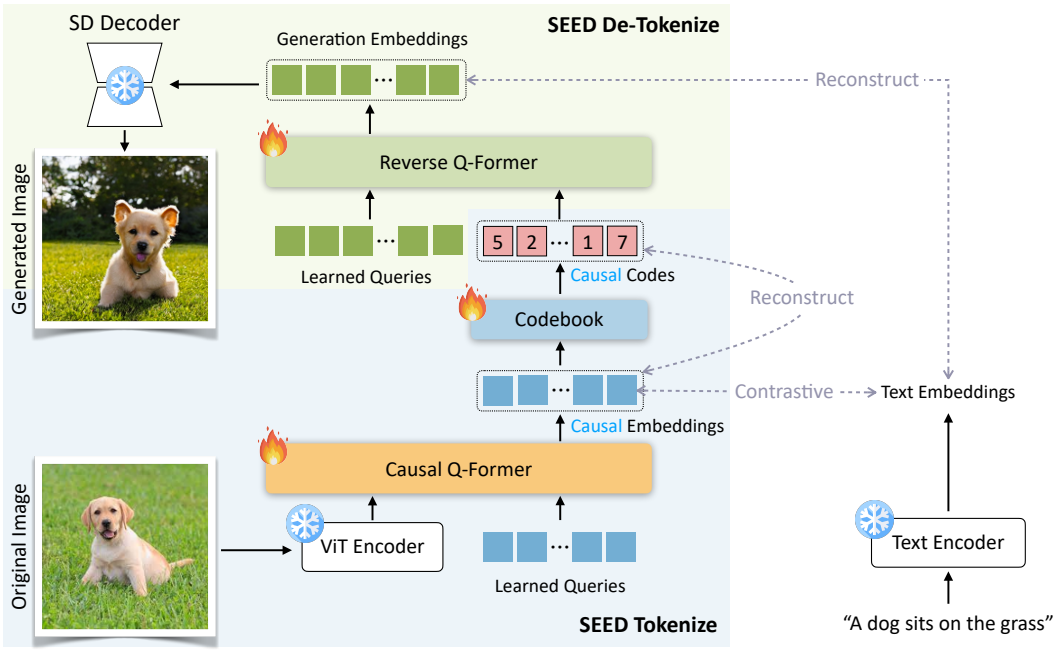

Figure 2: Overview of our **SEED** tokenizer, which produces discrete visual codes with causal dependency and high-level semantics.

Table 1: Evaluation of zero-shot Image-Text Retrieval. Causal codes are quantized causal embeddings.

Model	Flickr30K (1K test set)							COCO (5K test set)						
	Image → Text			Text → Image				Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	R@mean	R@1	R@5	R@10	R@1	R@5	R@10	R@mean
BLIP-2 [5]	81.9	98.4	99.7	82.4	96.5	98.4	92.9	65.3	89.9	95.3	59.1	82.7	89.4	80.3
SEED (causal emb)	90.0	99.6	99.9	80.0	95.3	97.6	93.7	71.9	91.1	95.9	56.7	80.7	87.7	80.7
SEED (causal code)	86.3	98.6	99.5	75.9	93.2	96.7	91.7	65.7	88.1	93.8	52.5	78.0	86.0	77.4

the **final** causal embedding and text features of the corresponding caption, while minimizing the similarity between the **final** causal embedding and text features of other captions in a batch.

Evaluation of Causal Embeddings. We evaluate the performance of Causal Q-Former on the zero-shot image-text retrieval task using **COCO** [21] and **Flickr30K** [24] dataset following BLIP-2. The performance is measured by *Recall@K* (R@K) for both image-to-text retrieval and text-to-image retrieval. Note that we adopt the dual-stream paradigm for inference and remove the image-txt-matching (ITM) rerank module in BLIP-2 for a fair comparison. As shown in Tab. 1, our Causal Q-former achieves better results than BLIP-2 in terms of an aggregated metric *Recall@mean*. It demonstrates that the output query embeddings with causal dependency do not drop performance than the output embeddings with bi-directional attention in BLIP-2.

2.2.2 Training Stage II: Visual Quantization and De-tokenization

As shown in Fig. 2, we train a VQ codebook to discretize the causal embeddings (32 tokens) into quantized visual codes (32 tokens) on 5M image-text pairs including CC3M, Unsplash, and COCO dataset. Specifically, a quantizer looks up the nearest neighbor in the codebook for each causal embedding and obtains the corresponding code. We employ a decoder, which is a multi-layer Transformer [22], to reconstruct the continuous causal embeddings from discrete codes. During training, we maximize the cosine similarity between the output of the decoder and the causal embeddings. We further employ a Reverse Q-Former to reconstruct the textual features of a frozen stable diffusion model from discrete codes. A set number of learnable query embeddings (77 tokens) are fed into the Reverse Q-Former. The query embeddings interact with each other through self-attention layers, and interact with causal codes (32 tokens) through cross-attention layers for the output generation embeddings (77 tokens). During training, we minimize the MSE loss between

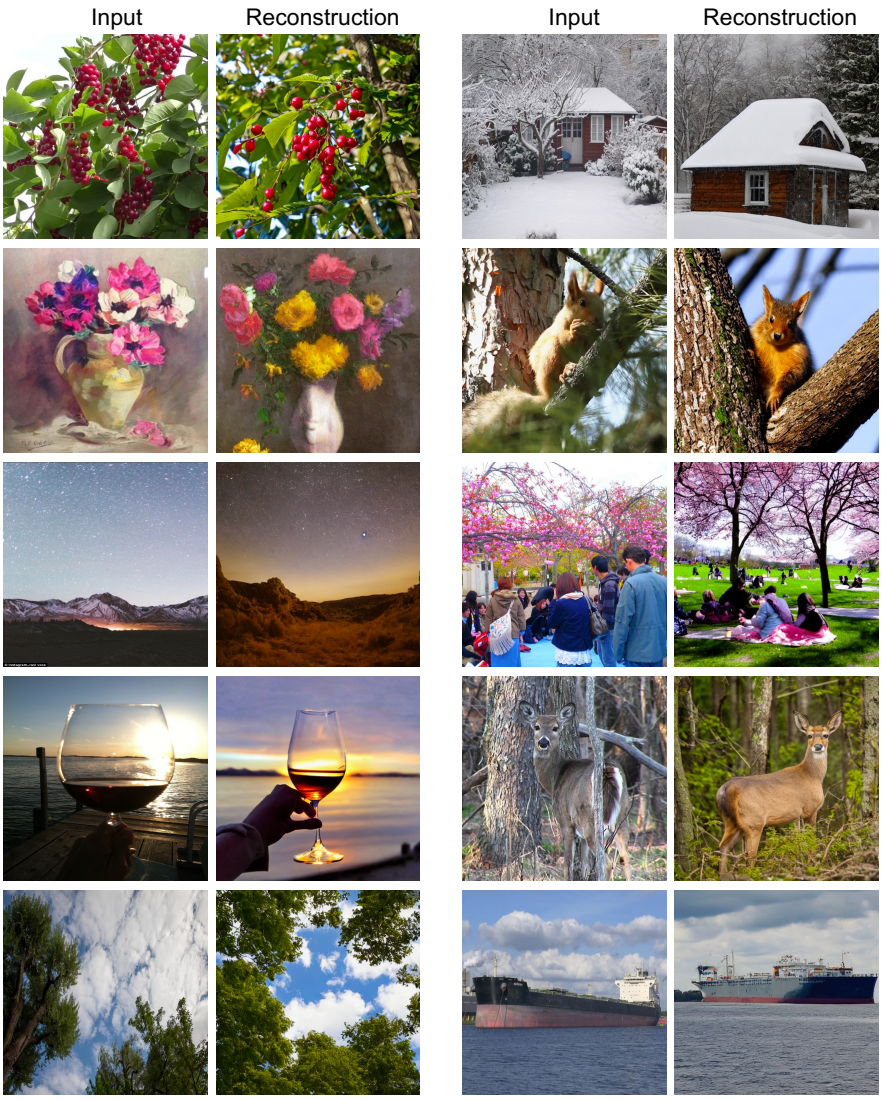

Figure 3: Reconstruction images of SEED tokenizer (i.e., original image → SEED tokenize → causal visual codes → SEED de-tokenize → reconstructed image), which are semantically consistent with the original input images.

generation embeddings and text features of SD. During inference, the generation embeddings can be fed into the SD-UNet to decode realistic images.

Evaluation of Causal Codes. We evaluate the performance of SEED tokenizer on zero-shot image-text retrieval, where the reconstructed causal embeddings from causal codes are used for retrieval. As shown in Tab. 1, discrete SEED tokens exhibit competitive performance compared to BLIP-2.

We further evaluate image generation on **COCO** and **Flickr30K** dataset. SEED first discretizes input images into causal codes (32 tokens) and obtain generation embeddings (77 tokens) from Reverse Q-Former, which are fed into the SD-UNet for the reconstructed images. For the baseline model GILL [25] and SD [10], images are generated from corresponding captions of the input images. We follow GILL [25] to compute the CLIP similarity as the evaluation metric for benchmarking the semantic consistency. As shown in Tab. 2, compared with the upper bound SD, our SEED only slightly drops performance, and outperforms GILL in image generation.

Table 2: Evaluation of Image Generation with CLIP similarity as the metric.

Model	COCO	Flickr30K
GILL [9]	67.45	65.16
SD [10]	68.43	65.40
SEED	68.23	65.22

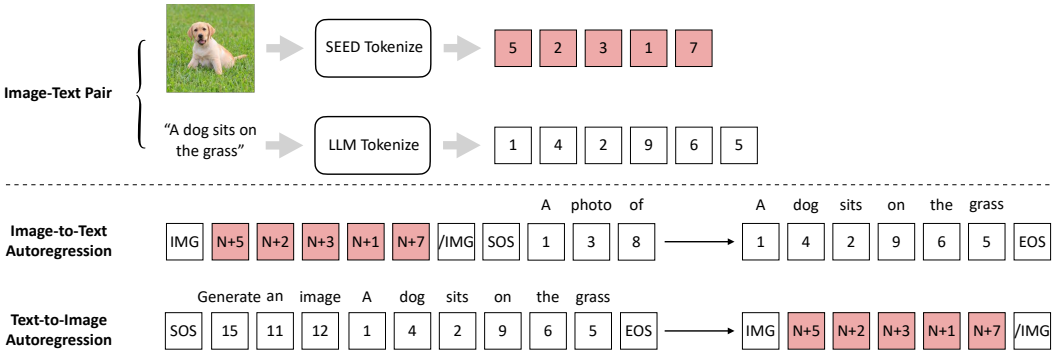

Figure 4: Overview of the multimodal autoregressive training for SEED-OPT$_{2.7B}$ using efficient LoRA tuning. It was trained in 44 hours using only 64 V100 GPUs and 5M image-caption pairs.

Visualization of Reconstructed Images. We visualize the reconstructed images of SEED in Fig. 3. Through utilizing the Reverse Q-Former to obtain the generation embeddings from the causal visual codes of the input image, realistic images can be generated using the off-the-shelf SD-UNet, which maintain consistent semantics with input images.

The above evaluation and visualization demonstrate the versatility of SEED visual tokens for both comprehension and generation tasks.

3 Multimodal Autoregression with SEED Visual Tokens

Based on the pre-trained SEED tokenizer, we present SEED-OPT$_{2.7B}$ through fine-tuning a low-rank adaption (LoRA) module on a OPT$_{2.7B}$ [19] model with 5M image-text pairs including CC3M, Unsplash and COCO dataset. As shown in Fig. 4, we perform image-to-text and text-to-image autoregressive pre-training for unified multimodal comprehension and generation.

Image-to-Text Autoregression. We first perform image-to-text autoregression to align the vocabulary of the pre-trained VQ codebook with OPT$_{2.7B}$. Specifically, we use a fully-connected (FC) layer to linearly project the causal codes from the visual tokenizer into the same dimension as the word embeddings of OPT$_{2.7B}$. The projected causal codes and the word embeddings of the prefix "A photo of" are concatenated as the input of the OPT$_{2.7B}$. The text tokens of the corresponding caption is used as the generation target. We freeze OPT$_{2.7B}$ and fine-tune LoRA with the training objective of predicting the next text token.

Text-to-Image Autoregression. We then jointly perform image-to-text and text-to-image autoregression to empower the LLM with the ability to generate vision tokens in addition to text tokens. For text-to-image autoregressive pre-training, the word embeddings of the prefix "Generate an image" and a caption are fed into OPT$_{2.7B}$. The visual codes of the corresponding image from our pre-trained tokenizer are used as the generation target. We freeze OPT$_{2.7B}$ and fine-tune LoRA with the training objective of predicting the next vision token.

During inference, given the prompt "Generate an image" and a text description, SEED-OPT$_{2.7B}$ predicts the visual tokens autoregressively. The output visual tokens are fed into the Reverse Q-Former for generation embeddings, which can be decoded to generate a realistic image via SD-UNet.

Evaluation of Multimodal Understanding. We evaluate the performance of SEED-OPT$_{2.7B}$ with zero-shot image captioning and visual question answering (vqa). For image captioning, we evaluate on both **COCO** [21] test set and **NoCaps** [26] validation set and report BLEU@K (B@K), METEOR (M), ROUGE$_L$ (R), CIDEr (C), and SPICE (S) with the prompt "a photo of". For visual question answering, we evaluate on **VQAv2** [27] validation set and **GQA** [28] test set and report Top-1 accuracy with the prompt "Question: {} Short answer." As shown in Tab. 3, compared with BLIP-2, which are trained on **129M** image-text pairs, our SEED-OPT$_{2.7B}$ trained on **5M** pairs achieves promising results on zero-shot image captioning and visual question answering with SEED discrete visual tokens. Note that different from concurrent work CM3Leon [12] that uses image captioning

Table 3: Comparison between BLIP-2 (pre-trained with 129M image-text pairs) and SEED-OPT$_{2.7B}$ (5M pairs) on zero-shot Image Captioning and Visual Question Answering. S: SPICE, M: METEOR, R: ROUGE$_L$, B: BLEU, C: CIDEr.

Models	NoCaps				COCO					VQAv2	GQA
	in S	near S	out S	overall S	Karpathy test M	R	B@4	C	S	Top-1	Top-1
BLIP-2 OPT$_{2.7B}$ [5]	14.4	13.8	13.4	13.8	39.7	28.9	59.3	131.0	22.9	51.9	32.6
SEED-OPT$_{2.7B}$	12.5	12.3	12.2	12.3	34.6	28.4	56.4	119.0	22.0	42.8	28.8

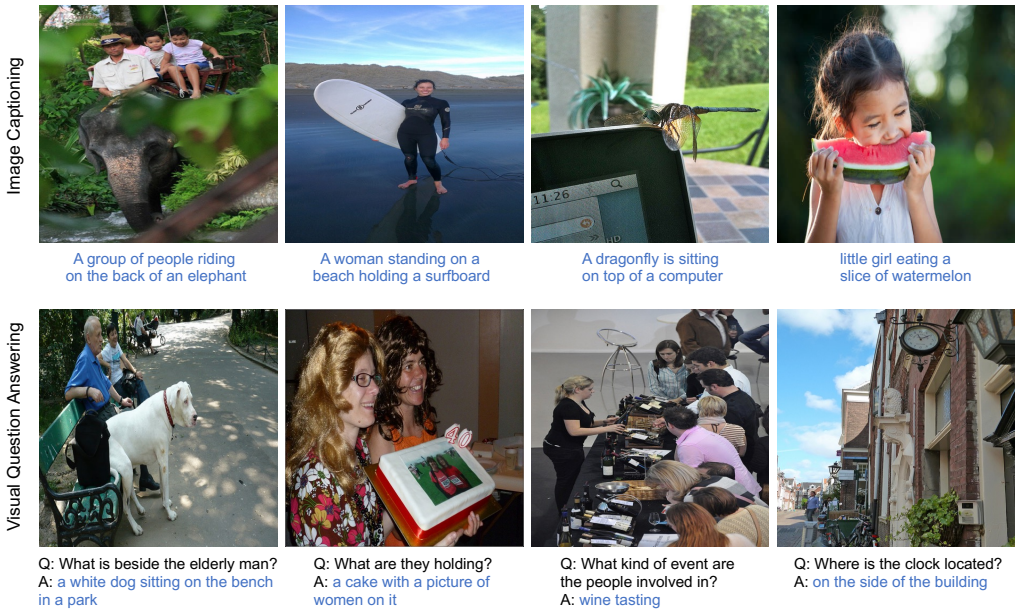

Figure 5: Qualitative examples of SEED-OPT$_{2.7B}$ on image captioning (with a prompt "a photo of") and open-ended visual question answering. Our model has not been trained on any VQA dataset.

and vqa datasets for **supervised fine-tuning**, our SEED-OPT$_{2.7B}$ pre-trained with image-to-text autoregression using the prefix "A photo of" can perform zero-shot visual question answering by understanding free-form questions and predicting open-form answers.

We also show qualitative examples of SEED-OPT$_{2.7B}$ on image captioning (with a prompt "a photo of") and vqa. As shown in Fig. 5, our model can generate captions than describe the visual content, and answer a variety of questions.

Evaluation of Multimodal Generation. We showcase qualitative examples of text-to-image generation results with our SEED-OPT$_{2.7B}$ in Fig. 6. Given the textual description, SEED-OPT$_{2.7B}$ can generate realistic images that are semantically relevant to the description.

SEED can facilitate alignment between visual tokens and LLMs, as evidenced by SEED-OPT$_{2.7B}$, already capable of performing text-to-image and image-to-text generation tasks after LoRA tuning.

4 Related Work

Multimodal Large Language Models for Comprehension. With the impressive success of Large language models [1, 2, 3] (LLMs), recent studies work on Multimodal LLM (MLLM) to improve visual comprehension through utilizing the strong generality of LLMs. Previous work [4, 5, 6, 29, 7, 8, 30, 31] align visual features of pre-trained image encoder with LLMs on image-text datasets, and empower LLMs with the ability to interpret visual information with textual descriptions. However, these work commonly use the prediction of the next text token as the training objective and exert no supervision for vision data, thus can only output texts given multimodal vision and language inputs.

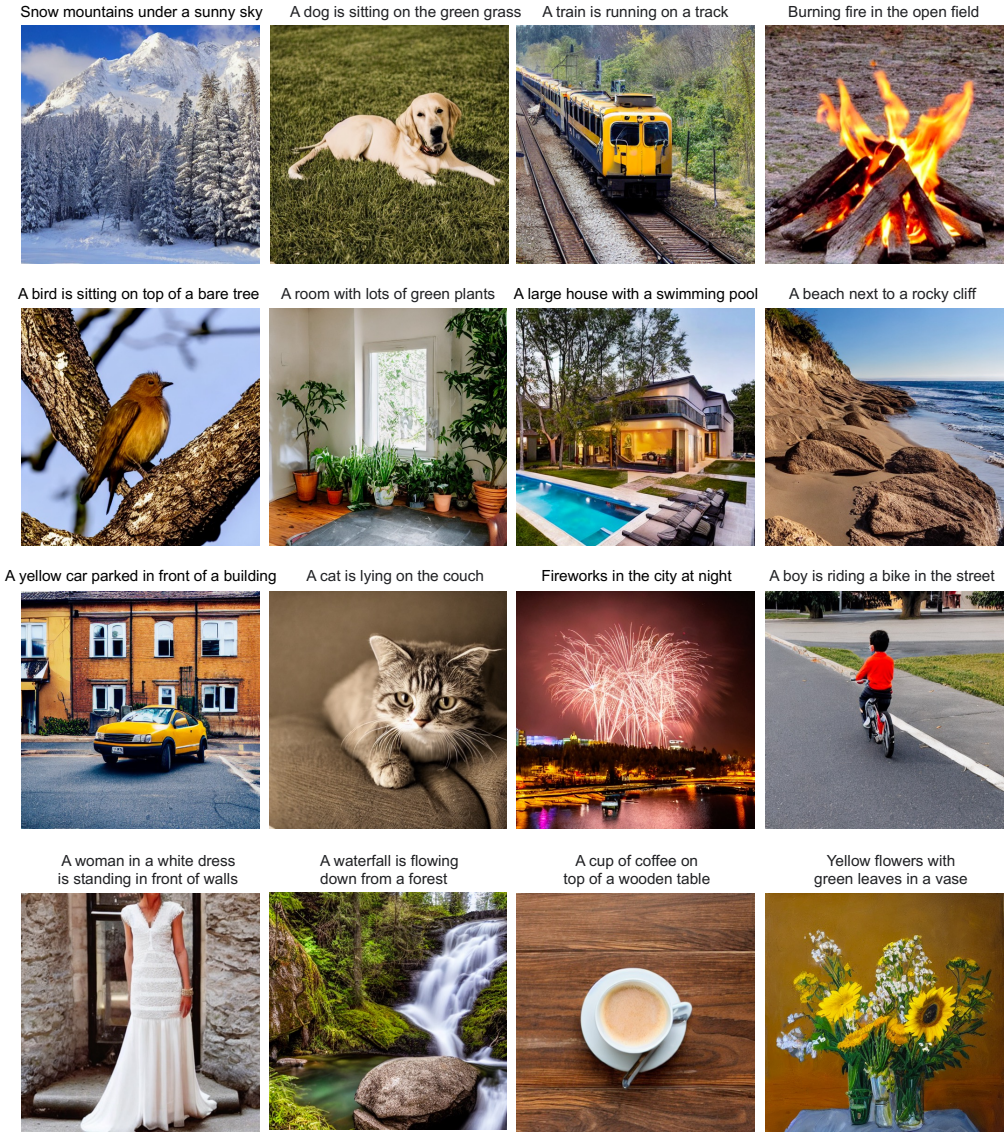

Figure 6: Text-to-image generation results when inferring with SEED-OPT$_{2.7\text{B}}$.

Multimodal Large Language Models for Generation. To empower LLMs with the image generation ability, CogView [14] pre-trains a visual tokenizer by reconstructing image pixels, and fine-tunes GPT models [2, 32] with the objective of next token prediction, where both image and text tokens are equally treated. GILL [25] learns a mapping between the embeddings of a LLM and a frozen pretrained image generation model. Both work aim to generate images with LLMs, without being explicitly designed for multimodal comprehension.

Visual Tokenizer. Visual tokenizer aims to represent the image as a sequence of discrete tokens similar to natural language. Previous work [15, 13, 17] trains a Vector Quantized Variational AutoEncoders (VQ-VAE) as a visual tokenizer by reconstructing the pixels of the input images, which captures only low-level details of images such as color, texture and edge. Beit v2 [18] trains a semantic-rich visual tokenizer through reconstructing high-level features from the teacher model, but its visual codes from 2D features of a vision transformer [22] are incompatible with the unidirectional attention in dominant LLMs for multimodal generation.

5 Conclusion

We present SEED, a discrete image tokenizer, designed based on the premise that visual tokens compatible with LLMs should capture high-level semantics while being generated with a 1D causal dependency. SEED enables LLMs to be trained with multimodal data following the original recipe of text (i.e., next-word prediction), which is mature and scalable. The trained multimodal LLM is capable of both image-to-text and text-to-image generation tasks, taking one more step toward emergent multimodal capabilities. We hope that our SEED would draw increased attention to visual tokenizers. A more rational visual tokenizer could substantially reduce the cost and complexity of multimodal LLM training, promoting lower-carbon, large-scale model training. Moreover, we eagerly anticipate the "germination" of vision (imagination) seeds within LLMs. The project is still in progress. Stay tuned for more updates!

Acknowledgements

We sincerely acknowledge Sijie Zhao (Tencent AI Lab) and Chen Li (ARC Lab, Tencent PCG) for their engaging discussions.

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] 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. *arXiv preprint arXiv:2204.02311*, 2022.

[4] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

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

[7] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[9] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

[10] 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.

[11] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[12] Yu Lili, Shi Bowen, Pasunuru Ram, Miller Benjamin, Golovneva Olga, Wang Tianlu, Babu Arun, Tang Binh, Karrer Brian, Sheynin Shelly, Ross Candace, Polyak Adam, Howes Russ, Sharma Vasu, Xu Jacob, Singer Uriel, Li (AI) Daniel, Ghosh Gargi, Taigman Yaniv, Fazel-Zarandi Maryam, Celikyilmaz Asli, Zettlemoyer Luke, and Aghajanyan Armen. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. 2023.

[13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[15] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[16] 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. 2023.

[17] 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.

[18] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[19] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[20] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Unsplash. https://github.com/unsplash/datasets.

[24] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[25] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.

[26] 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.

[27] 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.

[28] 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.

[29] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[30] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[31] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.