Pre-trained Language Models Do Not Help Auto-regressive Text-to-Image Generation

Yuhui Zhang[1], Brandon McKinzie[2], Zhe Gan[2], Vaishaal Shankar[2], Alexander Toshev[2]
[1]Stanford University, [2]Apple ML Research
yuhuiz@stanford.edu, {bmckinzie, z_gan, vaishaal_shankar, toshev}@apple.com

Abstract

Recent advances in image tokenizers, such as VQ-VAE, have enabled text-to-image generation using auto-regressive methods, similar to language modeling. However, these methods have yet to leverage pre-trained language models, despite their adaptability to various downstream tasks. In this work, we explore this gap by adapting a pre-trained language model for auto-regressive text-to-image generation, and find that pre-trained language models offer limited help. We provide a two-fold explanation by analyzing tokens from each modality. First, we demonstrate that image tokens possess significantly different semantics compared to text tokens, rendering pre-trained language models no more effective in modeling them than randomly initialized ones. Second, the text tokens in the image-text datasets are too simple compared to normal language model pre-training data, which causes the catastrophic degradation of language models' capability.

1 Introduction

Recent works in text-to-image generation primarily employ two kinds of methods: diffusion models [20, 25, 24] and auto-regressive models [21, 30]. The latter is facilitated by "image tokenizers", such as VQ-VAE [27, 23] and VQ-GAN [7, 29], which transform an image into a sequence of discrete tokens, similar to text tokens (Figure 1 Left). Consequently, image and text tokens can be jointly modeled using auto-regressive algorithms like the Transformer [28] (Figure 1 Right).

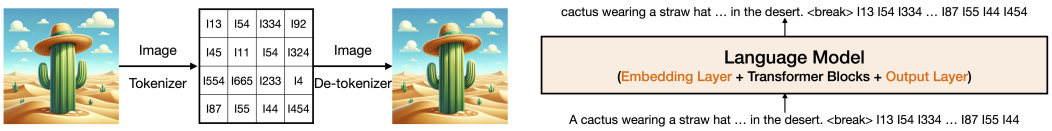

Figure 1: *Adapting language models for auto-regressive text-to-image generation. (Left)* An image is fed into an image tokenizer (MoVQGAN [36]) and converted to a grid of discrete tokens, and it can be well-reconstructed with these image tokens. *(Right)* As images are converted to tokens similar to text tokens, we can enable language models to generate images by adapting its embedding layer and output layer.

The superiority of diffusion-based models when compared with auto-regressive-based methods for text-to-image generation still remains unclear. OpenAI's pioneering work, DALL-E [21], showcased the potential of auto-regressive methods in this domain. Yet, its successor, DALL-E 2 [20], transitioned to a diffusion-based architecture and achieved enhanced image generation quality. Later, Google released Imagen [25] (diffusion-based) and Parti [30] (auto-regressive-based) at the same time and demonstrated their comparable generation quality. Similarly, the retrieval-augmented methods, Re-Imagen [4] (diffusion-based) and CM3leon [32] (auto-regressive-based), display similar

Work done during an internship at Apple ML Research.

NeurIPS ICBINB Workshop 2023.

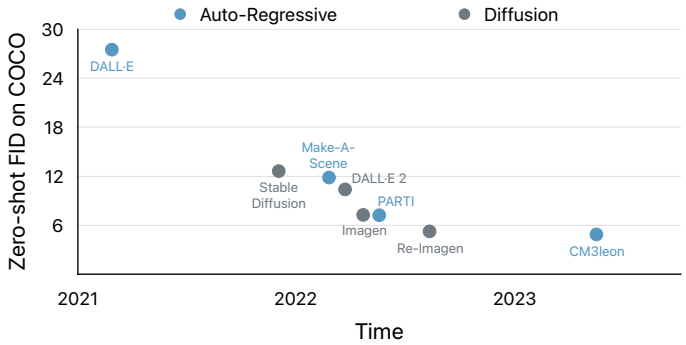

Figure 2: *Auto-regressive and diffusion based models achieve similar performances on text-to-image generation.* However, while all the diffusion models leverage pre-trained language models, all the auto-regressive models do not.

performance in text-to-image generation tasks. A comparison based on zero-shot FID [13] on the COCO dataset [15] can be found in Figure 2.

While these two approaches achieve similar performance, it is intriguing that *diffusion-based models consistently utilize pre-trained text encoders, whereas their auto-regressive counterparts generally do not*. For instance, Imagen [25] (diffusion-based) reports that employing a stronger pre-trained text encoder, specifically T5 [19], yields substantial improvements to using CLIP [18]. Furthermore, they observe that scaling up the T5 text encoder leads to more pronounced improvements than scaling up the diffusion models. Conversely, Parti [30] (auto-regressive-based) shows that using a pre-trained text encoder does not necessarily improve image quality in its Appendix. However, Parti employs an encoder-decoder architecture and uses BERT [6], a relatively inferior text encoder, to initialize the encoder only. It remains unclear whether a decoder-only approach would benefit from recent advances in large language models (LLMs), given the clear similarity between language modeling and auto-regressive text-to-image generation.

In this work, we explore the potential of pre-trained LLMs for auto-regressive text-to-image generation. To enable the model to process both text and image tokens, we expand the size of the embedding and output layers by incorporating an image vocabulary from the image tokenizer. We initialize these added weights either randomly or using a novel contrastive alignment (elaborated later in Section 3.2), while the remaining weights are directly copied from the original models. Subsequently, we fine-tune the model on image-caption datasets, as depicted in Figure 1 Right.

Surprisingly, the results show that pre-trained language models achieve the same loss and image generation quality as the model that is entirely randomly initialized and trained from scratch (Figure 3). Furthermore, we observe a catastrophic deterioration in the model's text capabilities, such as world knowledge or in-context learning, after only minimal steps of fine-tuning (Table 1).

To understand this phenomenon, we break down the cross-entropy loss on image and text tokens, and find that 1) the loss on image tokens is the same between the pre-trained and randomly initialized model, and 2) the loss on text tokens of the pre-trained model is significantly lower at the beginning compared to the randomly initialized models, but the gap soon disappears after training (Figure 4).

The first finding of the loss on the image tokens is particularly interesting. We hypothesize that image tokens obtained from image tokenizers might either lack semantics or possess significantly different semantics compared to text tokens, which renders language pre-training not transferable to the image modeling task. To verify this hypothesis, we conduct unconditional image generation experiments by training the model on image tokens only. Our results show that 1) the pre-trained model achieves the same loss as the randomly initialized model, and 2) freezing any part of the pre-trained model results in a loss degradation (Figure 5). These indicate that optimal weights for language and image modeling are fundamentally different, making language pre-training not transferable to image modeling.

In summary, we share our experimental findings about pre-trained language models do not help auto-regressive text-to-image generation, and offer an explanation: 1) the intrinsic differences between image and text tokens make language pre-training ineffective for the image token modeling, and 2) the disproportionate ratio between image and text tokens (usually 30:1 for image-caption datasets) minimizes the impact of loss on text tokens and leads to catastrophic forgetting.

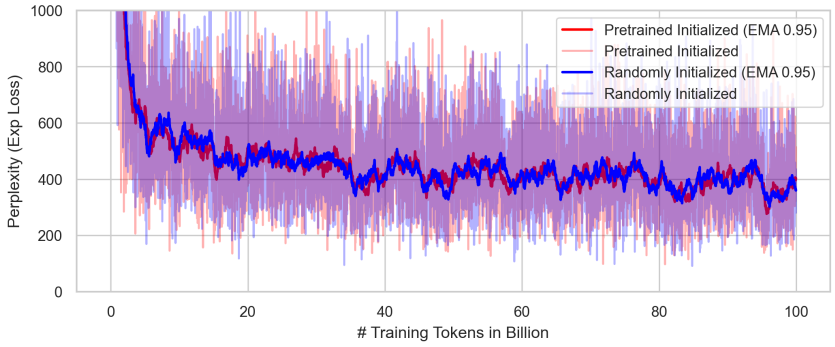

Figure 3: *Pre-trained language models do not help auto-regressive text-to-image generation.* Models are trained on the HQITP-134M image-caption dataset with 64 A100 80GB GPUs using batch size 1M tokens. EMA is Exponential Moving Average.

2 Pre-trained Language Models Do Not Help Text-to-Image Generation

2.1 Experimental Setup

Language model. We use the publicly available `open_lm` codebase and its `open_lm-1b` model for our experiments [10]. This language model contains ∼1B parameters and is trained on 1.6T tokens on a mix of RedPajama [5], Pile [8], S2ORC [16], The Pile of Law [11], Deepmind Math [26], and RealNews [34]. It achieves better or comparable performance compared to models with similar size such as OPT-1.3B [35], Pythia-1B [2], Neox-1.3B [3], OPT-IML-1.3B [14] on an average of 11 tasks such as HellaSwag [33] and MMLU [12]. More details can be found in the `open_lm` repository [10].

Image tokenizer. We use SBER-MoVQGAN [36] as the image tokenizer, which is the current state-of-the-art publicly available image tokenizer that achieves 0.686 FID on Imagenet image reconstruction. Given an image with 256 $\times$ 256 resolution, it converts an image to 1,024 tokens with a vocabulary size of 16,384. Figure 1 (Left) shows a real reconstruction example from this tokenizer.

Dataset. For multi-modal training, we use an internal dataset referred to as High Quality Image-Text Pairs (HQITP) [22], which contains 134M high-quality image-caption pairs. We pre-process the dataset before training. Each image is center-cropped to 256 $\times$ 256 and converted to 1,024 tokens. Each caption is tokenized with NeoX tokenizer with an average of 30 tokens. We add six special tokens corresponding to the beginning and end of document, text segment, and image, respectively. This results in input sequences of the form "<doc> <text> ...text tokens... </text> <image> ...image tokens... </image> </doc>", and pad them into 1,152 tokens with the special <pad> token.

Training setups. Models are trained with 100B tokens using 64 A100 80GB GPUs with batch size 1M tokens. We use the AdamW [17] optimizer with a cosine learning rate schedule with 2K warm-up steps and a peak learning rate of 0.0003. This mimics the settings reported in [1]. We also tried different hyperparameters, such as learning rates from 0.00005 to 0.0003 and batch size from 0.5M to 2M tokens, and found no significant influences on the conclusions.

2.2 Results

In Figure 3, we present the perplexity (exponential of loss) during training for both the pre-trained and randomly initialized models. Intriguingly, across the entire 100B token training regimen, the loss of the pre-trained model aligns closely with that of the randomly initialized one. Beyond this, a sharp decline in text capabilities of the pre-trained model is observed after training on 5B tokens, as illustrated in Table 1. At this point, both the model's world knowledge and its in-context learning ability are entirely diminished.

To delve deeper into this phenomenon, we separate the cross-entropy loss into two components: text tokens and image tokens, displayed separately in Figure 4. As anticipated, the pre-trained model begins with a significantly lower text loss in comparison to its randomly initialized counterpart. Yet, due to the overwhelming image-text token ratio (30:1), this initial advantage is obscured in the

Original Completion	Completion after Training 5B Tokens
Simply put, the theory of relativity states that **the speed of light is the same for all observers, regardless of their location in the universe.**	Simply put, the theory of relativity states that **iles must be able to see the invisible.**
Translate English to French: sea otter => loutre de mer peppermint => menthe poivrée plush girafe => girafe peluche cheese => **fromage**	Translate English to French: sea otter => loutre de mer peppermint => menthe poivrée plush girafe => girafe peluche cheese => **I love cheese**

Table 1: *Concrete examples of forgetting.* We observe a severe deterioration of the model's language capability, such as knowledge and in-context learning, after a small amount of training. Completions from the models are bolded.

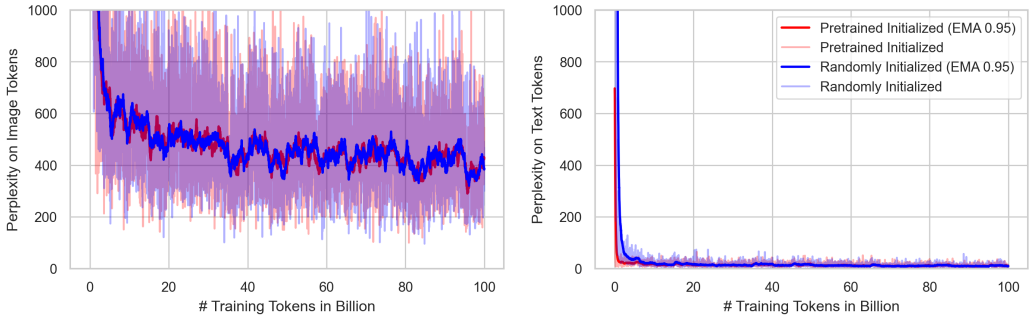

Figure 4: *Break-down loss on image and text tokens.* Models are trained on the HQITP-134M image-caption dataset with 64 A100 80GB GPUs using batch size 1M tokens.

aggregate loss. Furthermore, any benefit the pre-trained model offers in text loss diminishes soon during training. In contrast, for image tokens, there is no difference between the pre-trained and randomly initialized models. We hypothesize that the inability of effectively transferring a pre-trained language model to image token modeling is caused by the distinction between image and text tokens.

Moreover, loss on text tokens is substantially lower than image tokens, and even lower than typical language models trained on text-only data. This is because texts in image-caption datasets such as HQITP are less complex than those in standard text-only pre-training corpora, which also explains the catastrophic degradation of the model's text capability.

3 Image Tokens Are Drastically Different From Text Tokens

Why there is no difference between the loss of pre-trained and randomly initialized models on the image tokens? We hypothesize image tokens are significantly different from text tokens, for example, they lack semantics or have drastically different semantics compared to text tokens, which makes the pre-trained language model not transferable to image token modeling. Our unconditional image generation and image-token alignment experiments verify this hypothesis.

3.1 Unconditional Image Generation

To assess if pre-trained language models benefit image tokens, we perform unconditional image generation experiments. Unlike the text-to-image generation setup, we removed all text tokens, leaving only the image tokens. This approach rigorously examines if image tokens benefit from pre-trained language models. As shown in Figure 5, pre-trained language models yield the same loss as models initialized randomly.

Additionally, we selectively tune components of the pre-trained models: 1) only the embedding and output layer; 2) 1 plus layer norm and positional embedding; and 3) 2 plus the first half of layers; 4) 2 plus the feed-forward layers (FFN). Figure 5 presents these loss metrics. The findings reveal that none of these configurations achieves as low a loss as a fully tunable model. This underscores the

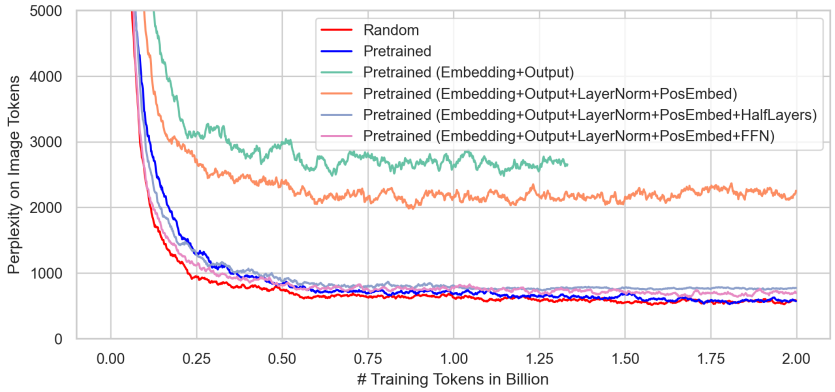

Figure 5: *Pre-trained language models do not help to model image tokens.* Models are trained only on the HQITP dataset's image tokens without any text tokens. We also compare the full fine-tuning with electively fine-tuning components of the pre-trained models (shown in parenthesis). EMA 0.95 is applied to the plot.

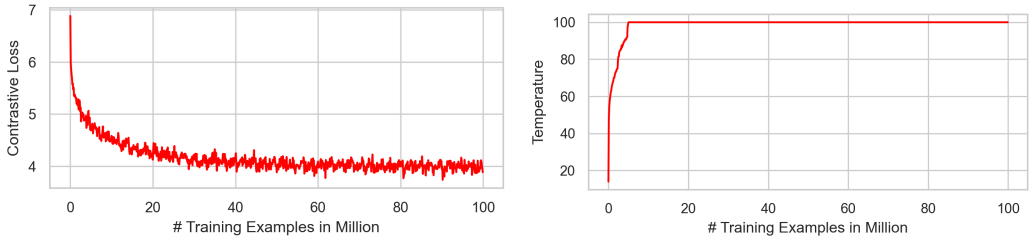

Figure 6: *Image-text token contrastive alignment. (Left)* The contrastive loss plateaus quickly, indicating a difficulty in aligning text and image tokens directly at a bag-of-words level. *(Right)* The learnable temperature in the contrastive loss during training for reference.

divergence in optimal weights for modeling text and image tokens, suggesting that any part of the text-trained weights is sub-optimal to transfer to image tokens.

3.2 Image-Text Token Contrastive Alignment

To understand whether image tokens have similar semantics as text tokens, we aligned image tokens with text tokens using a contrastive approach, inspired by methods like CLIP [18]. Given an image, we tokenize it into 1024 tokens and compute its bag-of-words image embeddings as its representation. Similarly, we tokenize the corresponding caption and compute its bag-of-words text embeddings. The text embeddings are initialized from a pre-trained language model while the image embeddings are randomly initialized. For a batch of $N = 1024$ image-caption pairs, the contrastive objective from CLIP is employed to maximize the cosine similarity between matched image-caption l_2-normalized representations and to minimize the similarity for non-matching pairs. Only the image embeddings are updated during training.

In Figure 6, we illustrate that the contrastive loss plateaus quickly, indicating a difficulty in aligning text and image tokens directly at a bag-of-words level. Indeed, after training, when querying the closest text tokens for any image token, we observe that they predominantly align with noisy, semantically void text tokens. Furthermore, when we use the trained image embeddings as initialization for text-to-image generation, as opposed to random initialization, there is no discernible improvement.

4 Conclusion

This study highlights the difficulty of naively adapting a text-only language model to handle multimodal contents, such as texts and images. Given the challenge of the disparities between image tokens and text tokens, a valuable avenue for future experiments is to employ tokenizers that align semantically with text tokens, such as SEED [9] or SPAE [31].

References

[1] A. Aghajanyan, L. Yu, A. Conneau, W.-N. Hsu, K. Hambardzumyan, S. Zhang, S. Roller, N. Goyal, O. Levy, and L. Zettlemoyer. Scaling laws for generative mixed-modal language models. In *ICML*, 2023.

[2] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.

[3] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *ACL Workshop*, 2022.

[4] W. Chen, H. Hu, C. Saharia, and W. W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *ICLR*, 2022.

[5] T. Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[7] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

[8] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[9] Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

[10] S. Gururangan, M. Wortsman, S. Y. Gadre, A. Dave, M. Kilian, W. Shi, J. Mercat, G. Smyrnis, G. Ilharco, M. Jordan, R. Heckel, A. Dimakis, A. Farhadi, V. Shankar, and L. Schmidt. open_lm: a minimal but performative language modeling (lm) repository, 2023. GitHub repository.

[11] P. Henderson, M. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *NeurIPS*, 2022.

[12] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[14] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[16] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The semantic scholar open research corpus. In *ACL*, 2020.

[17] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[20] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[21] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[22] K. Ranasinghe, B. McKinzie, S. Ravi, Y. Yang, A. Toshev, and J. Shlens. Perceptual grouping in contrastive vision-language models. In *ICCV*, 2023.

[23] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[25] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[26] D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. In *ICLR*, 2019.

[27] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[29] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved VQGAN. In *ICLR*, 2022.

[30] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022.

[31] L. Yu, Y. Cheng, Z. Wang, V. Kumar, W. Macherey, Y. Huang, D. A. Ross, I. Essa, Y. Bisk, M.-H. Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. *arXiv preprint arXiv:2306.17842*, 2023.

[32] L. Yu, B. Shi, R. Pasunuru, B. Muller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023.

[33] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In *ACL*, 2019.

[34] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. In *NeurIPS*, 2019.

[35] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[36] C. Zheng, T.-L. Vuong, J. Cai, and D. Phung. Movq: Modulating quantized vectors for high-fidelity image generation. In *NeurIPS*, 2022.