Enhancing Large Multi-Modal Auto-Regressive Models with Condition Contrastive Alignment

Chendong Xiang
Department of Computer Science
2024316344
xiangxyaw@gmail.com

Mingdao Liu
Department of Computer Science
2024316083
joshua720117@gmail.com

Yuji Wang
Department of Computer Science
2024310693
wangyuji24@mails.tsinghua.edu.cn

Abstract

The rapid development of auto-regressive (AR) models in multi-modal generation has brought promising advancements, enabling coherent text, image, and video generation within a single framework. However, AR models still face significant challenges in practical application, especially in image generation where classifier-free guidance (CFG) is commonly used to enhance output quality. CFG, while effective, introduces substantial computational overhead and deviates from the simplicity of end-to-end auto-regressive generation. In this proposal, we aim to explore the potential of Condition Contrastive Alignment (CCA) within Emu3, a state-of-the-art multi-modal AR model, to address the reliance on CFG in image generation. By applying CCA, a recently proposed method for aligning AR models with target distributions through contrastive learning, we hypothesize that Emu3 can achieve comparable or superior output quality without CFG, reducing computational cost and improving generation efficiency. Our approach involves fine-tuning Emu3 with CCA on multi-modal data and conducting comprehensive evaluations across image and video generation benchmarks. This research will validate CCA's applicability to large AR models, potentially advancing the field towards more efficient, unified multi-modal generation frameworks.

1 Background and Related Works

Multi-modal Models. Multi-modal models [10, 13, 18, 21] have advanced rapidly in recent years, driven by the latest breakthroughs in language and vision models, particularly auto-regressive (AR) language models [1, 14] and diffusion-based visual generative models [2, 8, 12]. This line of research aims to develop models capable of handling multi-modal generation (*e.g.* text-to-image generation, text-to-video generation) and perception (*e.g.* vision-language understanding) tasks within a single framework. In this project, we focus on AR multi-modal models [19, 20], which are considered to have considerable potential due to the simplicity and scalability of auto-regressive methods. These models unify text, image and video data into discrete tokens, training and inference with the *next-token prediction* approach.

Guided Sampling. Although the training and inference of language and vision data can be unified through *next-token prediction* with auto-regressive models, there is still a gap in the sampling process of auto-regressive language and vision models. To enhance sample quality, visual generative

auto-regressive models rely on guided sampling methods [3, 6, 10, 12], which adjust the *sampling distribution* by modifying the sampling algorithm without fine-tuning the pre-trained model. Specifically, classifier-free guidance (CFG) masks the condition with a relatively low rate (*e.g.* 10%) during training, enabling the model to predict unconditional logits. Then, a combination of conditional and unconditional logits is used when sampling [9, 11]. CFG complicates the original training method of auto-regressive models (*i.e. next-token prediction*), and double the computational overhead of sampling. In contrast, auto-regressive language models leverage alignment fine-tuning based on Reinforcement Learning from Human Feedback (RLHF) to improve instruction-following abilities by adjusting the *model distribution* and keeping the sampling algorithm unchanged [1, 15]. Recently, Condition Contrastive Alignment (CCA) has been proposed to guide the sampling of visual generative auto-regressive models through a fine-tuning algorithm [4] derived from Noise Contrastive Estimation (NCE), providing an approach to unifying the sampling of auto-regressive language and vision models.

2 Proposed Method

Problem Formulation. Consider a sample (*e.g.* an image) $\boldsymbol{x}$ represented by a sequence of N discrete tokens $\boldsymbol{x} = \{x_1, x_2, ..., x_N\}$. The probability of sample $\boldsymbol{x}$ given condition $\boldsymbol{c}$ (*e.g.* the description of the image) can be decomposed as:

$$p(\boldsymbol{x}|\boldsymbol{c}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{x}_{<n}, \boldsymbol{c}) \tag{1}$$

Each token $\boldsymbol{x}_n$ is conditioned only on $\boldsymbol{c}$, which can also be represented as discrete tokens, and its previous input $\boldsymbol{x}_{<n}$. An auto-regressive (AR) model θ learns the conditional probability $p_\theta(\boldsymbol{x}_n|\boldsymbol{x}_{<n}, \boldsymbol{c})$ and samples tokens one by one in generation.

Review of CCA Method. To enhance the sample quality under condition $\boldsymbol{c}$, CCA [4] derives a fine-tuning method from guided sampling method and Noise Contrastive Estimation (NCE), where the loss is defined as

$$\mathcal{L}_\theta^{\text{CCA}} = -\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{c})} \log \sigma\Big[\frac{1}{s} \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}\Big] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{c})} \log \sigma\Big[-\frac{1}{s} \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{c})}{p_\phi(\boldsymbol{x}|\boldsymbol{c})}\Big]. \tag{2}$$

p_ϕ is a pre-trained AR model and is frozen during training. p_θ is the target model and is initialized from p_ϕ. s is the guidance scale. Intuitively, $\mathcal{L}_\theta^{\text{CCA}}$ maximizes the relative likelihood where the condition $\boldsymbol{c}$ and the sample $\boldsymbol{x}$ matches, and minimize the the relative likelihood where the condition $\boldsymbol{c}$ and the sample $\boldsymbol{x}$ are independent and most likely mismatch.

Review of practical CCA Method. To tractably sample from the joint distribution $p(\boldsymbol{x}, \boldsymbol{c})$ and the product of two independent marginals $p(\boldsymbol{x})p(\boldsymbol{c})$, CCA [4] propose a practical training loss. Consider a training batch with K samples $B = \{(\boldsymbol{x}_k, \boldsymbol{c}_k)^{1:K}\}$. A random permutation of $\boldsymbol{c}_k$ in B is used as samples from $p(\boldsymbol{x})p(\boldsymbol{c})$ and the original batch B as samples from $p(\boldsymbol{x}, \boldsymbol{c})$. Then the loss for fine-tuning is defined as

$$\mathcal{L}_\theta^{\text{CCA}} = -\log \sigma\Big[\beta \log \frac{p_\theta(\boldsymbol{x}_k|\boldsymbol{c}_k)}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_k)}\Big] - \lambda \log \sigma\Big[-\beta \log \frac{p_\theta(\boldsymbol{x}_k|\boldsymbol{c}_{\rho(k)})}{p_\phi(\boldsymbol{x}_k|\boldsymbol{c}_{\rho(k)})}\Big] \tag{3}$$

where ρ is a random permutation, β and λ are two hyper-parameters.

Research Plan. We plan to fine-tune Emu3 [20] with CCA method on multi-modal data and conduct comprehensive evaluations across visual generation benchmarks. Our preliminary proposal is to use JourneyDB [17], a dataset with over 4 million high-resolution images and corresponding annotations, as the fine-tuning dataset and to evaluate the generation performance with Fréchet Inception Distance [7] (FID) and Inception Score [16] (IS) on ImageNet [5] dataset. Considering the image resolution, fine-tuning iterations and model size, we estimate the fine-tuning process will require approximately one week on 8 NVIDIA A100 GPUs. This research will explore CCA's applicability to large AR models and more complex tasks, potentially advancing the field towards more efficient, unified multi-modal generation frameworks.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[4] Huayu Chen, Hang Su, Peize Sun, and Jun Zhu. Toward guidance-free ar visual generation via condition contrastive alignment. *arXiv preprint arXiv:2410.09347*, 2024.

[5] 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.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[11] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[12] 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.

[13] 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.

[14] 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.

[15] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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

[18] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[20] 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.

[21] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.